USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Trabajo de Fin de Grado

Revisitando los modelos ANOVA y ANCOVA

Alejandro Arias Piñeiro

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo de Fin de Grado

Revisitando los modelos ANOVA y ANCOVA

Alejandro Arias Piñeiro

Julio 2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e investigación operativa.

Título: Revisitando los modelos ANOVA y ANCOVA

Breve descripción del contenido

Los modelos de análisis de la varianza (ANOVA, del inglés Analysis of Variance) y análisis de la covarianza (ANCOVA, del inglés Analysis of Covariance) se pueden ver como casos particulares del modelo lineal general, y en sus casos más simples, se considera una variable explicativa de tipo factor (modelo ANOVA) o bien esta variable está acompañada de una o varias variables escalares (modelo ANCOVA). Las técnicas de inferencia sobre estos modelos no están exentas de hipótesis (normalidad, independencia de observaciones, homogeneidad de varianzas). Sin embargo, en la práctica se suelen utilizar estos modelos sin realizar una validación adecuada de las hipótesis y sin tener en cuenta las consecuencias de que dichas hipótesis no se cumplan.

En este trabajo se revisará la inferencia sobre los modelos ANOVA y ANCOVA, y mediante un exhaustivo estudio de simulación, se analizará el impacto de no cumplir alguna (o varias) de las hipótesis habituales.

Para ilustrar los efectos del incumplimiento de las hipótesis, se revisará la literatura en algún campo aplicado, tratando de identificar ejemplos de mal uso de estos modelos.

Recomendaciones

Es recomendable que el alumno tenga capacidad para manejar bibliografía en inglés. En la parte de programación, será necesario programar código en R.

Índice general

Resumen

El objetivo de este trabajo será analizar que ocurre cuando fallan las hipótesis del modelo ANOVA y ANCOVA, las cuales son la homogeneidad de varianzas, la normalidad y la independencia de observaciones. Con este objetivo en mente, primero introduciremos los modelos y las herramientas que utilizaremos en el estudio de simulación, para posteriormente, ponerlo en práctica.

Así comenzaremos a evaluar en términos de calibrado y potencia el comportamiento del test F en los escenarios en los que se cumplan una mayor cantidad de hipótesis. Posteriormente, iremos suprimiendo estás hipótesis y compararemos los resultados obtenidos con los resultados de los escenarios originales. Finalmente, extraeremos conclusiones del estudio de simulación de los modelos ANOVA y ANCOVA.

Abstract

The objective of this dissertation is to analyze what happens when the hypothesis of the ANOVA and ANCOVA models fail. These hypotheses are variance homogeneity, normality and independence of the observations. With this goal on mind, we will introduce first the models and the statistic tool that we will use in the simulation study. Secondly, we will show the performance in practice.

Thus, we will begin to evaluate the behaviour of the F test in terms of calibration and power in the scenarios, fulfilling totally the required hypothesis. Then, some scenarios relaxing this assumptions are considered, and results will be compared with the original scenario. From an extensive simulation study, we will draw some conclusions.

Introducción

En una gran cantidad de estudios médicos es necesaria la comparación de la media de dos o más grupos, como, por ejemplo, cuando se desea testar la eficacia de un nuevo medicamento para una enfermedad. Generalmente, para este propósito, se emplea un **modelo ANOVA**. Este modelo se construye bajo tres hipótesis: *normalidad*, *independencia de observaciones* y *homogeneidad de varianzas*.

En esta clase de estudios es de vital importancia realizar una correcta validación del modelo y cumplir las hipótesis. En caso de que no se cumplan, podemos llegar a falsas conclusiones, como, por ejemplo, que un nuevo medicamento es eficaz ante una enfermedad cuando realmente no existen pruebas para apoyar esta hipótesis.

Por desgracia, a pesar de que el modelo ANOVA es ampliamente requerido, en numerosas investigaciones no se verifica si los datos bajo estudio cumplen las hipótesis para el buen funcionamiento de este en la práctica. Un ejemplo es el artículo de Wu et al. (2011) donde se pone de manifiesto que este es un problema recurrente:

> Apparently, due to the researcher's lack of basic knowledge of statistics, they ignored the application condition of a certain method. When the quantitative data did not meet the prerequisites for parametric tests, they blindly applied the tests.

En la asignatura **Modelos de Regresión y Análisis Multivariante**, estudiamos el modelo ANOVA y aprendemos a validar sus hipótesis. Sin embargo, en muchos casos, no somos conscientes del impacto que tendría la ausencia de alguna de estas y no profundizamos en ninguna alternativa a este modelo en caso de que no se cumplan dichas hipótesis.

En este trabajo presentaremos el contraste de igualdad de medias en el modelo ANOVA y veremos como se comporta (en términos de calibrado y potencia) cuando falla alguna de las hipótesis habituales. Además, pese a no ser una de las hipótesis propias del modelo, estudiaremos el efecto que un diseño muestral desbalanceado puede tener sobre el método

clásico. Con este propósito, realizaremos un estudio de simulación en el que haremos uso de las **técnicas de Monte Carlo**: analizando el calibrado y la potencia del test a través del porcentaje de rechazos (empíricos) cuando los datos se generan bajo la hipótesis nula de igualdad de medias o bajo alguna alternativa, quebrantando una o varias de las hipótesis que sustentan las labores de inferencia sobre el modelo.

No obstante, existen situaciones donde en el contraste de igualdad de medias entra en juego una segunda variable continua que aporta información relevante sobre cada uno de los grupos considerados. Por ejemplo, existen medicamentos cuyo efecto sobre un paciente dependen de sus características fisiológicas, como puede ser el peso o la edad. De esta forma, si se quiere probar la eficacia de un nuevo medicamento de esta gama, tendremos que tener en cuenta esta segunda variable continua a la hora de llevar a cabo el contraste de igualdad de medias, ya que puede no ser necesaria la misma dosis para una persona con sobrepeso que para una persona delgada. En este contexto surge el **modelo ANCOVA**.

Este modelo será analizado en la segunda parte del trabajo y se construye nuevamente bajo las hipótesis de *normalidad*, *independencia de observaciones* y *homogeneidad de varianzas*. De nuevo es importante la correcta validación de las hipótesis del modelo, usando test como el **test de Shapiro-Wilks** o el **test de Levene**. En caso de no llevarse a cabo, se pueden comprometer gravemente las conclusiones extraídas del estudio.

Para poner este modelo a prueba, realizaremos un estudio similar al del modelo ANOVA, pero con una particularidad: en el modelo ANCOVA tendremos dos contrastes de hipótesis diferentes y habrá que analizarlos por separado. Esto es debido a que ahora, a diferencia del modelo ANOVA, tendremos dos variables explicativas diferentes y habrá que contrastarlas individualmente.

Para ello realizaremos los **Test de no efecto** de la variable continua y discreta. Con el objetivo de comprobar el no efecto de estás variables, fue creado un código de simulación en el cual haremos de nuevo uso de las **técnicas de Monte Carlo**.

Finalmente, sacaremos conclusiones sobre los resultados obtenidos y compararemos el funcionamiento del test F en los diferentes escenarios. En el caso del modelo ANOVA también propondremos brevemente alguna alternativa al test F.

Capítulo 1

Introducción de los modelos ANOVA y ANCOVA

En este capítulo introduciremos los modelos ANOVA y ANCOVA enmarcándolos en el contexto del modelo lineal general, para, a posteriori, analizar la necesidad de las distintas hipótesis supuestas a la hora de realizar inferencia.

Para llevar esto a cabo, en la Sección 1.1 presentaremos los diferentes modelos de regresión y los encasillaremos en el modelo lineal general. En la Sección 1.2 veremos más a fondo el modelo ANOVA, presentando la notación y las suposiciones básicas. La Sección 1.3 la dedicaremos a ver si el papel del grupo aporta información sobre la variable respuesta Y. Más tarde en las Secciones 1.4 y 1.5 haremos lo propio con el modelo ANCOVA.

1.1. Breve revisión del modelo lineal general

Empezaremos recordando que se entiende por **modelo de regresión**. En muchos problemas existe relación entre dos o más variables, una variable de interés, a la que denominaremos *variable respuesta* Y y un conjunto de variables con las que trataremos de explicar la variable respuesta, a las que llamaremos *variables explicativas* $X = (X_1, \ldots, X_{p-1})$. Esto puede ser escrito como:

$$Y = f(X) + \varepsilon, \tag{1.1}$$

donde en (1.1) f es una función de X desconocida y ε es el término del error, el cual es independiente de X y tiene media cero.

El error ε incluye todo aquello que no podemos cuantificar y por lo tanto no podemos

usar para predecir Y. Además, incluye todas las fuentes de variación no medibles. Por ejemplo, el riesgo a una reacción adversa puede variar en un paciente determinado y en un día determinado, dependiendo de cómo haya sido fabricado el medicamento o como se sienta ese día el paciente (James et al. (2013)).

Ejemplo 1.1. Un ejemplo muy común es el estudio de Galton (Moore (2005)) de la dependencia de la estatura de los hijos (Y) respecto a la de sus padres (X). En este se encontró una relación entre ambas: los padres altos tienen en general hijos altos, aunque de media no tan altos como sus padres, mientras que los padres bajos tienen hijos bajos, aunque de media más altos que sus padres. Pues bien, un modelo matemático que explique la relación entre estas variables será un modelo de regresión.

Cuando suponemos que en nuestro modelo de regresión la función f es lineal, diremos que es un **modelo de regresión lineal**. Ahora, la estructura del modelo (1.1) vendría dada por la expresión:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon, \tag{1.2}$$

siendo $\beta_0, \ldots, \beta_{p-1}$ constantes que hay que estimar, conocidas como parámetros de la función de regresión. Es importante mencionar que el método utilizado normalmente para estimar $\beta_0, \ldots, \beta_{p-1}$ es el de mínimos cuadrados, el cual trata de minimizar la suma de diferencias al cuadrado entre el valor observado y el valor predicho. Además, en este escenario tanto la variable respuesta Y como el vector de variables explicativas X son cuantitativas (toman valores numéricos) y continuas.

En este modelo se considera, por lo general, que el error ε satisface las hipótesis de:

- **Homocedasticidad**: la varianza del error es la misma para cualquier valor de la variable explicativa X.
- **Normalidad**: el error tiene distribución normal

$$\varepsilon \in N(0, \sigma^2)$$

- **Independencia**: conocida la formulación del modelo lineal (1.2), si tenemos una muestra de n elementos, debemos suponer que las variables aleatorias que representan los errores $\varepsilon_1, \ldots, \varepsilon_n$ son mutuamente independientes.

Estas se pueden formular como:

$$\varepsilon_1, \ldots, \varepsilon_n \in N(0, \sigma^2) \text{ y son independientes.}$$

El modelo de regresión lineal puede ser expresado en términos del **modelo lineal general** (véase McCullagh (2018)) como:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

el cual se puede expresar mediante la notación abreviada:

$$Y = X\beta + \varepsilon, \tag{1.3}$$

donde ahora en (1.3) Y es el vector de respuestas, $X \in \mathbb{R}^{n\times p}$ es una matriz que contiene la información relativa a los individuos por filas y en cada columna se muestra una cierta variable explicativa, por β denotamos al vector de parámetros y ε es un vector que contiene los errores y verifica $\varepsilon \in N(0, \sigma^2 I_n)$. El modelo lineal general incluye varios modelos estadísticos diferentes como ANOVA, ANCOVA o regresión lineal simple entre otros.

1.2. Modelo ANOVA

A diferencia de los modelos de regresión lineal, en el **modelo de análisis de la varianza** o **ANOVA**(Maxwell et al. (2017)), hay una única variable explicativa X y es discreta, mientras que la variable respuesta Y sigue siendo continua. Que la variable explicativa X sea discreta es un gran cambio, ya que mientras en el Ejemplo 1.1 la altura de los padres podía ser cualquier valor numérico, ahora X toma la forma de una categoría, como veremos en el Ejemplo 1.2.

Ejemplo 1.2. Trataremos de estudiar la eficacia de diferentes tratamientos contra la anorexia. La base de datos se denomina `anorexia` y se encuentra en el paquete `MASS` del software `R` (R Core Team (2019)). Incluye datos de 72 mujeres que padecieron este trastorno, referidos al tratamiento usado con ellas (Cont hace referencia al grupo de control, CBT significa que fueron tratadas con un tratamiento cognitivo conductual y FT que fueron tratadas con un tratamiento familiar) y el incremento o disminución de peso entre el inicio y el final del periodo de estudio.

Como vemos en la Figura 1.1 en este caso la variable explicativa X son los diferentes tratamientos usados y es una variable discreta. Es una simple etiqueta identificativa del

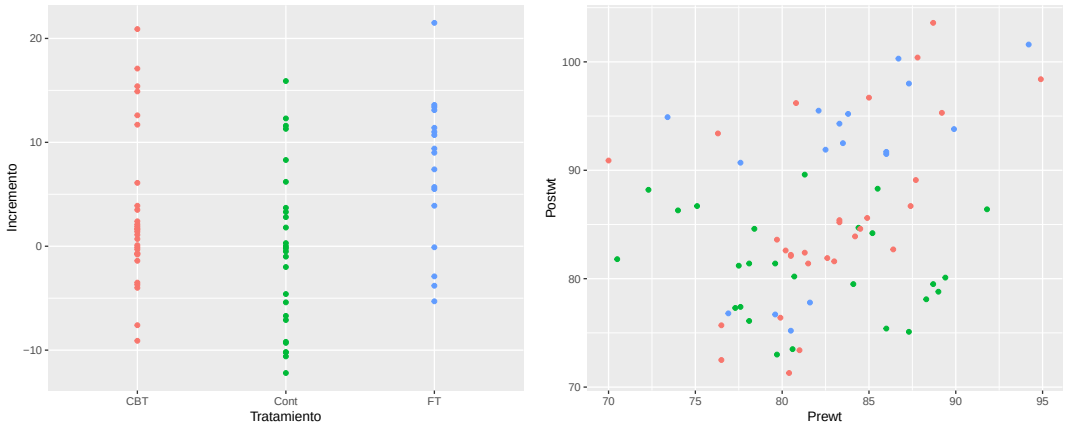

Figura 1.1: Izquierda: diagrama de dispersión del incremento de peso en libras de las mujeres en función del tratamiento usado. Derecha: diagrama de dispersión del peso después del tratamiento sobre el peso antes del tratamiento. Los diferentes tratamientos en los individuos se representan mediante colores, rojo (CBT), verde (control) y azul (tratamiento familiar).

tratamiento. Por otro lado, la variable respuesta Y es la diferencia de peso en libras de las mujeres después del tratamiento. En este contexto tiene sentido la comparación de los diferentes tratamientos.

Como acabamos de ver en el Ejemplo 1.2, la variable explicativa produce la descomposición de la muestra en I submuestras con n_i elementos cada una, $i = 1, \ldots, I$, donde cada submuestra procede de una población diferente. Estas observaciones se pueden denotar así:

$$\begin{array}{llll} Y_{11} & Y_{12} & \ldots & Y_{1n_1} \quad \text{de una población } N(\mu_1, \sigma^2) \\ Y_{21} & Y_{22} & \ldots & Y_{2n_2} \quad \text{de una población } N(\mu_2, \sigma^2) \\ & & \ldots & \ldots \quad \ldots \\ Y_{I1} & Y_{I2} & \ldots & Y_{In_I} \quad \text{de una población } N(\mu_I, \sigma^2) \end{array}$$

En este contexto seguimos suponiendo que las muestras son independientes y siguen todas una distribución normal dentro de cada grupo. Cabe destacar que a las medias de cada grupo se les permite ser distintas, pero las varianzas se asumen todas iguales.

La idea en este modelo es aprovechar la información que aporta el grupo al que pertenece una observación para decir algo de ella. Más específicamente, dada una variable explicativa discreta X y una variable respuesta continua Y, construiremos un modelo de manera que

dependiendo de a cual de las I poblaciones pertenezca nuestro individuo, podamos dar una mejor estimación de nuestra variable respuesta Y. Esta estimación será la media del grupo al que pertenece nuestra observación, es decir:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad i \in \{1, \dots, I\} \quad j \in \{1, \dots, n_i\} \tag{1.4}$$

donde los $\varepsilon_{ij} \in N(0, \sigma^2)$ son independientes y μ_i es un vector de parámetros que contiene la media de cada población.

Como en la Sección 1.1, el modelo debe de satisfacer una serie de hipótesis:

- **Homocedasticidad** [1]: La varianza de las I poblaciones es la misma.
- **Normalidad**: Las observaciones de cada grupo i provienen de una distribución Normal.
- **Independencia**: Las I muestras son, entre sí, independientes.

Estas hipótesis serán necesarias a la hora de realizar inferencia sobre la población. En el Capítulo 2 veremos que ocurre a la hora de realizar un estudio ante la ausencia de estas hipótesis.

Además, mediante notación matricial, el modelo (1.4) se puede expresar en términos del modelo lineal general (1.3) como

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & \cdots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & 0 & 1 \\ \vdots & \cdots & \cdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{I1} \\ \vdots \\ \varepsilon_{In_I} \end{pmatrix}$$

[1] Cuando usamos variables discretas, nos referimos a la igualdad de varianzas como homogeneidad de varianzas. Sin embargo, en este trabajo usaremos el término empleado en el caso de que la variable sea continua. De igual forma emplearemos el término **heterocedástico** en lugar de heterogeneidad de varianzas.

De este modo, el modelo de Análisis de la Varianza queda planteado como un modelo lineal general.

1.3. El test F en el modelo ANOVA

Dado que el objetivo es aprovechar la información de cada grupo $i \in \{1, \dots, I\}$, primero es necesario plantearse si realmente la discretización en grupos, mediante la variable X, nos aporta información importante sobre la variable respuesta Y. Es decir, si existen diferencias significativas entre los grupos o si, por el contrario, el papel de grupo no aporta información relevante. Para comprobar esto, se lleva a cabo un contraste de hipótesis. De esta forma, daremos por cierta la igualdad de medias $\mu_1 = \cdots = \mu_I$ y veremos si la muestra aporta pruebas en su contra. Por lo tanto, nuestras hipótesis son:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$

$$H_a : \exists\ k, l \text{ tal que } \mu_k \neq \mu_l$$

La primera de ellas, H_0, es denominada hipótesis nula, mientras que la segunda es conocida como hipótesis alternativa. Su contraste es un problema de decisión, que podemos representar mediante la siguiente tabla:

Realidad / Decisión	Aceptar	Rechazar
H_0 cierta	Correcto	Error de tipo I
H_0 falsa	Error de tipo II	Correcto

Observemos que podemos tomar una decisión correcta o errónea. Llamaremos **error de tipo I** al que cometemos cuando rechazamos H_0 siendo esta cierta, mientras que el **error tipo II** será el que cometemos cuando aceptamos H_0 siendo esta falsa. Definiremos ahora el nivel de significación y la potencia:

- **Nivel de significación**: Es la probabilidad de error tipo I, y lo denotaremos por α.
- **Potencia**: Es la probabilidad, cuando H_0 es falsa, de rechazarla, la cual se corresponde con $1 -$ Error tipo II. A la probabilidad de error tipo II la denotaremos por ω.

Un test que acepte siempre H_0 no cometerá nunca un error tipo I, sin embargo, el error tipo II será muy alto, por lo que el test no tendrá un comportamiento adecuado. La idea es encontrar el equilibrio entre ambas, para ello fijaremos un nivel de significación y nos

quedaremos con el criterio que nos proporcione la mayor potencia posible. Este criterio suele estar basado en un estadístico de contraste, que refleja si los datos son más compatibles con H_0 o H_a.

En nuestro caso, compararemos la cantidad de variabilidad explicada por nuestro modelo con la cantidad de variabilidad no explicada. Para verlo con detenimiento, presentaremos RSS_0 como la suma residual de cuadrados bajo la hipótesis nula mientras que RSS será la suma residual de cuadrados bajo la alternativa:

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \qquad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2 \tag{1.5}$$

donde

$$\bar{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \qquad \forall i \in \{1, \ldots, I\} \tag{1.6}$$

es la media local del grupo i y

$$\bar{Y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{I} n_i \bar{Y}_{i\bullet} \tag{1.7}$$

es la media global. Los elementos de las ecuaciones (1.5) admiten la siguiente interpretación:

- RSS_0: Representa la variabilidad total del modelo, tomando las diferencias entre cada una de las observaciones y la media global.
- RSS: Representa la variabilidad dentro de cada uno de los grupos, siendo la media del grupo i la mejor estimación que se puede dar para ese grupo. Mide la diferencia entre la estimación y la observación, es decir, recoge lo que el modelo no es capaz de explicar.
- $RSS_0 - RSS$: Es la variabilidad total menos la variabilidad no explicada, o lo que es lo mismo, la variabilidad que el modelo es capaz de explicar.

Hemos dicho que queremos emplear un estadístico que compare la cantidad de variabilidad explicada por nuestro modelo con la cantidad de variabilidad no explicada. Haciendo uso de la descomposición de RSS_0 dada por:

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \tag{1.8}$$

se obtiene la siguiente descomposición de la variabilidad que se muestra en la Tabla 1.1.

Fuente de variación	Suma de cuadrados	Grados de libertad
Entre grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2$	$I-1$
Dentro de los grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2$	$n-I$
Total	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2$	$n-1$

Tabla 1.1: Tabla de análisis de la varianza.

De esta forma, para comparar la variabilidad explicada con la no explicada, se obtiene el estadístico

$$F=\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2/(n-I)}=\frac{(RSS_0-RSS)/(I-1)}{RSS/(n-I)}, \tag{1.9}$$

el cual da lugar al **test** $\boldsymbol{F}$.

Sin embargo, nos queda saber cual es la distribución de este estadístico, pues en ella se basarán los puntos de corte o de rechazo. La distribución del cociente de varianzas recibe el nombre de F de Snédecor, y la definimos a continuación.

Definición 1.3. Si $X_1 \in \chi^2_{m_1}$, $X_2 \in \chi^2_{m_2}$ y son independientes entonces

$$F=\frac{X_1/m_1}{X_2/m_2}\in F_{m_1,m_2}$$

y decimos que F tiene distribución F de Snedecor con m_1 grados de libertad en el numerador y m_2 grados de libertad en el denominador.

Como los errores ε_{ij} son normales e independientes con media cero y varianza σ^2, cada Y_{ij} proviene de una población normal $N(\mu_i,\sigma^2)$ y los grados de libertad del numerador y denominador suman $n-1$ como se puede ver en la Tabla 1.1. Entonces, tanto el numerador como el denominador siguen una distribución chi cuadrado χ^2 y son independientes. Por lo tanto, el test F (1.9) sigue una distribución F de Snédecor:

$$F=\frac{(RSS_0-RSS)/(I-1)}{RSS/(n-I)}\in F_{(I-1),(n-I)} \tag{1.10}$$

Por consiguiente, fijado un nivel de significación $\alpha \in (0,1)$, rechazaremos H_0 cuando el estadístico F (1.9) sea mayor de f_α, siendo este el cuantil que deja a su derecha una probabilidad α en la distribución $F_{(I-1),(n-I)}$. Como podemos ver, el estadístico es más grande cuanto mayor es la diferencia entre $\bar{Y}_{i\bullet}$ y $\bar{Y}_{\bullet\bullet}$ con $i = 1, \ldots, I$ y más pequeño cuanto mayor es la diferencia entre Y_{ij} y $\bar{Y}_{i\bullet}$ con $i = 1, \ldots, I$, $j = 1, \ldots, n_i$.

Esto tiene sentido, pues $\bar{Y}_{i\bullet}$ con $i = 1, \ldots, I$ es la media local estimada de la población i, por lo que cuanto mayor sea la diferencia de esta con $\bar{Y}_{\bullet\bullet}$, la media global estimada, más fácil es que el papel del grupo aporte información.

Por el contrario, cuanto mayor es la diferencia entre Y_{ij} (cada observación) y $\bar{Y}_{i\bullet}$ (la media del grupo de esa observación) con $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, mayor será la variabilidad de esa población, por lo que es más fácil confundir un efecto del grupo con un efecto aleatorio.

1.4. Modelo ANCOVA

Como comentamos al final de la Sección 1.1, el modelo lineal general también incluye el **modelo ANCOVA**(Maxwell et al. (2017)). En este modelo, intervienen al mismo tiempo variables explicativas que son discretas y continuas. En este trabajo, en el escenario del modelo ANCOVA, consideraremos que X está formado por una variable discreta y por una variable continua. Para ilustrarlo seguiremos con el Ejemplo 1.4.

Ejemplo 1.4. Continuamos usando la base de datos `anorexia` procedente del paquete `MASS` del software `R`. En este caso, además de los diferentes tratamientos, incluiremos el peso anterior al tratamiento y el peso después del tratamiento en el modelo.

En la imagen derecha de la Figura 1.1 podemos observar que, por norma general, a mayores valores de peso antes del tratamiento le corresponden mayores valores de peso tras el tratamiento. Aún así, como podemos ver en la figura, parece que no hay una relación lineal muy clara. Además, también parece que que hay diferencia entre los distintos tratamientos usados y el grupo de control.

En este ejemplo las variables explicativas serían el tratamiento usado y el peso antes del tratamiento, mientras que la variable respuesta sería el peso después del tratamiento. En algunos casos podría tener sentido usar efectos de interacción, ya que un determinado tratamiento podría ser más eficaz en casos agudos de anorexia, por ejemplo.

Por lo tanto, en el contexto anterior, estamos considerando una situación donde conocemos la variable respuesta de cada individuo Y_{ij}, el grupo $i = 1, \dots I$ y al mismo tiempo, otra variable explicativa continua que podemos representar como z_{ij}. Como consecuencia, podemos considerar un modelo de regresión que incluya el efecto de ambas variables

$$Y_{ij} = \mu + \eta_i + \gamma z_{ij} + \varepsilon_{ij} \quad i \in \{1, \dots, I\} \quad j \in \{1, \dots, n_i\} \tag{1.11}$$

donde μ representa una constante, η_i el efecto del grupo i y γ el coeficiente de regresión de la variable continua z. Como en los anteriores casos, se supone que los $\varepsilon_{ij} \in N(0, \sigma^2)$ son independientes.

Por otro lado, las variables podrían interactuar entre sí. Esto llevaría a que una misma variación de z afectaría de forma diferente a la variable respuesta Y en función del grupo al que pertenece, por lo que en este caso, tendríamos rectas de regresión con diferente pendiente. Esto lo podríamos parametrizar como:

$$Y_{ij} = \mu + \eta_i + \gamma z_{ij} + \delta_i z_{ij} + \varepsilon_{ij} \quad i \in \{1, \dots, I\} \quad j \in \{1, \dots, n_i\}$$

donde δ_i sería la variación en las pendientes de las rectas de regresión en cada grupo.

Ejemplo 1.5. La base de datos se denomina `whiteside` procedentes del paquete `MASS` del software `R`. Incluye datos de viviendas del Reino Unido referidos al consumo de gas semanal, a la temperatura media exterior e información acerca de si las viviendas poseen aislamiento o no.

En este caso las variables explicativas X son la información acerca del aislamiento y la temperatura exterior, mientras que la variable respuesta Y es el consumo de gas semanal. Parece que, como podemos ver en la Figura 1.2, la relación lineal es más clara, a la vez que los grupos están mejor definidos.

En este contexto podríamos preguntarnos si sería adecuado ajustar dos rectas de regresión con diferente pendiente para cada grupo o si, por el contrario, dos rectas de regresión con la misma pendiente sería suficiente.

1.5. El test F en el modelo ANCOVA

A diferencia del modelo ANOVA, no solo debemos plantearnos aprovechar la información de cada grupo $i \in \{1, \dots, I\}$, sino que debemos contrastar el efecto de la variable

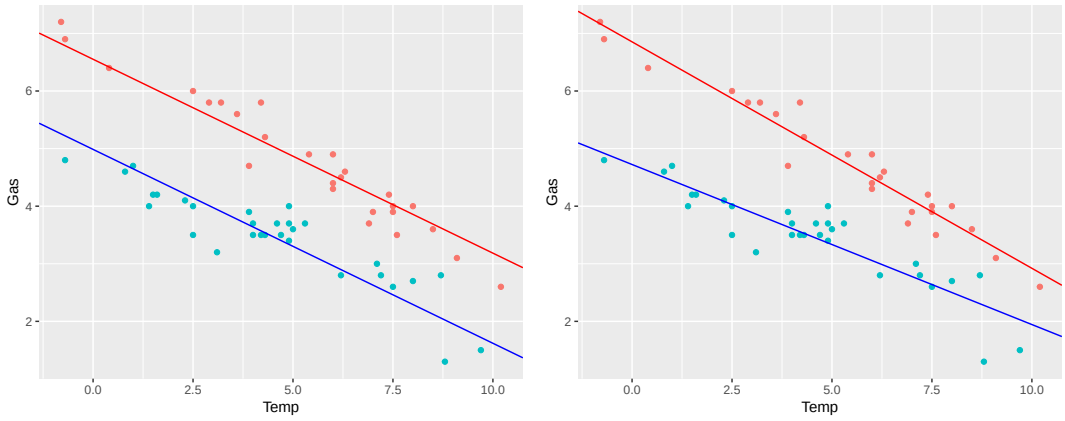

Figura 1.2: Izquierda: diagrama de dispersión del consumo de gas en función de la temperatura exterior, rectas de regresión ajustadas con la misma pendiente en cada grupo. Derecha: diagrama de dispersión del consumo de gas en función de la temperatura exterior, rectas de regresión ajustadas con diferente pendiente en cada grupo. El color rojo representa a las casas sin aislamiento, mientras que el color azul pertenece a las casas con él.

continua. Por ello, primero nos plantearemos el contraste del efecto de esta variable y por último, estudiaremos el impacto de la variable discreta.

1.5.1. Contraste del efecto de la variable continua

Para llevar esto a cabo, realizaremos un contraste de hipótesis en el que daremos por cierto el no efecto de la variable continua y veremos si la muestra aporta pruebas en su contra. Nuestras hipótesis son:

$$H_0 : \gamma = 0$$

$$H_a : \gamma \neq 0$$

donde γ es el coeficiente de regresión de la variable z en el modelo (1.11). Para realizar el contraste, emplearemos de nuevo el test F. Recordemos de la Sección 1.3 que este está basado en las sumas residuales del modelo bajo H_0 y bajo H_a. Así, si H_0 es cierta, estamos en un modelo usual de análisis de la varianza, cuya suma residual sería

$$RSS_0 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 \tag{1.12}$$

mientras que

$$RSS = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu} - \hat{\eta}_i - \hat{\gamma} z_{ij})^2 \tag{1.13}$$

donde $\bar{Y}_{i\bullet}$ es la media local del grupo i (1.6). Como consecuencia, el test F estaría basado en el siguiente estadístico y distribución:

$$F = \frac{(RSS_0 - RSS)}{RSS/(n - I - 1)} \in F_{1,n-I-1} \tag{1.14}$$

donde RSS_0 y RSS son los presentados en las ecuaciones (1.12) y (1.13) respectivamente.

1.5.2. Contraste del efecto de la variable discreta

En este caso, nos plantearemos si es necesario incluir la variable discreta en el modelo o si, por el contrario, no tiene ningún efecto sobre Y. El contraste de hipótesis será el siguiente:

$$H_0 : \eta_i = 0 \ \forall i$$

$$H_a : \exists \eta_i \text{ tal que } \eta_i \neq 0$$

Por lo tanto, si la hipótesis nula fuese cierta, estaríamos en un modelo de regresión lineal simple sobre z, siendo la suma residual de cuadrados bajo H_0

$$RSS_0 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet} - \hat{\gamma}_0 (z_{ij} - \bar{z}_{\bullet\bullet}))^2 \tag{1.15}$$

donde $\bar{Y}_{\bullet\bullet}$ es la media global de Y (1.7) y $\bar{z}_{\bullet\bullet}$ es la media global de la variable continua z. El RSS sería el mismo que el mostrado en la ecuación (1.13). Por último, el estadístico adoptaría la forma

$$F = \frac{(RSS_0 - RSS)/(I - 1)}{RSS/(n - I - 1)} \in F_{I-1,n-I-1}. \tag{1.16}$$

Tanto en el contraste de la variable continua como en el contraste de la variable discreta, fijado un nivel de significación α, rechazaremos H_0 cuando el valor de los estadísticos de las ecuaciones (1.14) y (1.16) sea mayor de f_α, siendo este el cuantil que deja a su derecha una probabilidad α en las distribuciones $F_{1,n-I-1}$ y $F_{I-1,n-I-1}$ respectivamente.

Capítulo 2

Estudios de simulación sobre el modelo ANOVA

En este capítulo pondremos a prueba el test F del modelo ANOVA, suprimiendo distintas hipótesis y viendo como responde el test.

Para realizar esta tarea, en la Sección 2.1 introduciremos el objetivo de esta capítulo. En la Sección 2.2 presentaremos los diferentes escenarios de simulación que tendremos en cuenta. La Sección 2.3 estará dedicada a contar el procedimiento que fue utilizado para realizar el estudio de simulación, mientras que la Sección 2.4 recoge el algoritmo usado para simular los datos. Posteriormente, en las Secciones 2.5 y 2.6 daremos e interpretaremos los resultados del estudio de simulación en términos del calibrado y de la potencia del test. La Sección 2.7 será empleada para extraer conclusiones del estudio realizado. Por último, en la Sección 2.8, hablaremos de otros test alternativos al test F.

2.1. Objetivo

Como hemos visto, tenemos una herramienta (el test F) que nos permite ver si el papel del grupo tiene algún efecto sobre la variable respuesta Y. Recordemos que la hipótesis nula y alternativa eran:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$

$$H_a : \exists\ k, l \text{ tal que } \mu_k \neq \mu_l.$$

A la hora de implementar un contraste,fijamos un nivel de significación $\alpha \in (0, 1)$,siendo la probabilidad de error tipo I (Tabla 1.3), y verificamos si hay pruebas significativas para

rechazar H_0. De esta forma, si los datos están bajo la hipótesis nula H_0, queremos equivocarnos a lo sumo un porcentaje α de las veces al rechazar esta. En consecuencia, si la probabilidad de rechazo del test está próxima al α fijado, diremos que el test está bien **calibrado**. Esto se puede comprobar mediante la simulación de datos bajo H_0, contando el porcentaje de veces que el test rechaza esta hipótesis. De la misma manera, es necesario verificar que un test no acepte siempre la hipótesis nula. Para ello, es suficiente simular datos bajo la hipótesis alternativa y comprobar qué proporción de veces es rechazada por el test. A la probabilidad de rechazar H_0 cuando estamos bajo H_a se denomina **potencia** del test.

El **objetivo** de este capítulo será evaluar el impacto de la ausencia de alguna de las hipótesis del modelo ANOVA. Para este fin, estudiaremos el comportamiento del test en términos de calibrado y potencia cuando alguna de las hipótesis falla. Por lo tanto:

- **Un test se comportará bien en términos de calibrado** cuando la probabilidad de rechazo del test F esté próxima al α fijado.
- **Un test se comportará bien en términos de potencia** cuando, en un mismo escenario, al ir aumentando el tamaño muestral la potencia se aproxima cada vez más a 1.

Por último, también tendremos en cuenta el **p-valor**, que es la probabilidad de obtener valores más grandes que el del estadístico observado. Si el test está bien calibrado, estos deberían de seguir una distribución Uniforme $U[0, 1]$.

2.2. Escenarios de simulación

En esta parte evaluaremos el test en términos de calibrado y potencia. A continuación presentamos 6 escenarios de simulación diferentes, en todos ellos se han considerado $I = 3$ clases:

1. **Diseño balanceado (E1)**: las desviaciones típicas de los grupos 1, 2 y 3 serán $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$ respectivamente. En lo que se refiere al tamaño de cada muestra en el estudio del calibrado, las tres serán del mismo tamaño $n_1 = n_2 = n_3 = 50$.
2. **Diseño desbalanceado (E2)**: las desviaciones típicas de los grupos 1, 2 y 3 serán $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$ respectivamente. En este caso, en el estudio del calibrado, el tamaño de muestra del primer grupo será $n_1 = 200$, mientras que el tamaño

del segundo y tercer grupo será $n_2 = n_3 = 50$ respectivamente. En este escenario observaremos qué ocurre cuando las muestras están desbalanceadas.

3. **Diseño con mayor variación desbalanceado (E3)**: las desviaciones típicas de los grupos 1, 2 y 3 serán $\sigma_1 = \sigma_2 = \sigma_3 = 1$ respectivamente. En este caso, en el estudio del calibrado, el tamaño de muestra del primer grupo será $n_1 = 100$, mientras que el tamaño del segundo será $n_2 = 50$ y el del tercero será $n_3 = 25$. En este escenario observaremos qué ocurre cuando las tres varianzas aumentan, a la par que nos enfrentaremos nuevamente a las muestras desbalanceadas.

4. **Diseño heterocedástico balanceado (E4)**: en lo que se refiere al estudio del calibrado, el tamaño de cada muestra será el mismo para las tres, tomando el valor $n_1 = n_2 = n_3 = 50$. En este escenario testearemos por primera vez qué ocurre cuando infringimos una de las hipótesis del modelo, la homocedasticidad. Dividiremos este en dos subescenarios:

   - **Diseño heterocedástico 1 (E4.1)**: las desviaciones típicas de los grupos 1 y 2 serán $\sigma_1 = \sigma_2 = 0{,}5$ respectivamente, mientras que la desviación típica del grupo 3 será $\sigma_3 = 2$.
   - **Diseño heterocedástico 2 (E4.2)**: las desviaciones típicas de los grupos 1 y 2 serán $\sigma_1 = \sigma_2 = 0{,}5$ respectivamente, mientras que la desviación típica del grupo 3 será $\sigma_3 = 0{,}1$. A diferencia del anterior escenario, veremos que ocurre cuando la desviación típica de uno de los grupos es menos que 0,05.

5. **Diseño heterocedástico desbalanceado (E5)**: las desviaciones típicas de los grupos 1 y 2 serán $\sigma_1 = \sigma_2 = 0{,}5$ respectivamente, mientras que la desviación típica del grupo 3 será $\sigma_3 = 0{,}1$. En este caso, en el estudio del calibrado, el tamaño de muestra del primer y segundo grupo serán $n_1 = n_2 = 50$ respectivamente, mientras que el tamaño del tercero será $n_3 = 200$.

6. **Diseño con errores no normales balanceado (E6)**: las muestras estarán balanceadas, siendo los tamaños de cada grupo en el estudio del calibrado $n_1 = n_2 = n_3 = 50$. Este será el primer escenario en el cuál se incumpla la hipótesis de normalidad. Lo dividiremos en dos subescenarios:

   - **Diseño con errores T de Student (E6.1)**: en este escenario, las observaciones de los grupos seguirán todas una distribución T de Student con 3 grados de libertad. En este escenario, a pesar de no cumplirse la hipótesis de normalidad, la distribución de los errores es simétrica.

- **Diseño con errores asimétricos (E6.2)**: en este escenario, las observaciones de los grupos seguirán todas una distribución χ^2 con 5 grados de libertad. En este subescenario la distribución de los errores es asimétrica.

Además, como el estudio se realiza en términos de calibrado y potencia, es necesario simular bajo H_0 para comprobar el comportamiento en términos de calibrado y bajo H_a para comprobar el comportamiento en términos de potencia. Por ello, se ha decidido tomar el valor $\mu = 1$ para la media de los grupos 1, 2 y 3 en el estudio del calibrado. Sin embargo, como el estudio de la potencia depende del tamaño muestral, veremos el comportamiento del test según aumente este.

2.3. Código y metodología

Para realizar este estudio de simulación, se ha empleado el lenguaje de programación *R* (R Core Team (2019)). En él, se ha simulado $M = 1000$ veces cada uno de los escenarios introducidos en la Sección 2.2.

Como mencionamos al comienzo, estudiaremos el comportamiento en términos de calibrado y potencia. Por ello, en el estudio del calibrado, fijaremos un $\alpha = 0{,}05$ y veremos cuál es la probabilidad de que el test rechace la hipótesis nula H_0. Para este fin se calcula, en media, el número de veces que se ha rechazado la hipótesis nula en las M=1000 simulaciones cuando los datos están bajo H_0. A esta probabilidad la denotaremos con la letra δ. Más tarde, en el estudio de la potencia, observaremos la probabilidad de que el test rechace la hipótesis nula H_0, pero en este caso simulando bajo H_a. Esta probabilidad será $\theta = 1 - \omega$. Además, diremos que el test F tiene un comportamiento bueno en términos de potencia si al ir aumentando el tamaño muestral la potencia se aproxima cada vez más a 1.

Es importante mencionar que la potencia del test variará dependiendo del **número de observaciones**, de la **varianza del error** y de la **diferencia en las medias de los distintos grupos**. Por ello, variaremos las diferentes hipótesis y veremos como cambia la potencia y el calibrado del test.

Dado que el objetivo es evaluar el impacto de la ausencia de alguna de las hipótesis del modelo ANOVA, comenzaremos con escenarios que cumplen la mayoría de estas y veremos como evoluciona el comportamiento del test F según vamos prescindiendo de hipótesis. Por lo tanto, tomaremos como grupo de referencia el Primer Escenario, también llamado

Diseño balanceado.

Observaremos como se comporta el test cuando se infringen dos de sus hipótesis, la homocedasticidad y la normalidad. Por ello en el cuarto y quinto escenario se simulará usando diferentes tamaños muestrales y desviaciones típicas en los grupos, mientras que en el sexto se simulará de manera que:

- Las observaciones proceden de una distribución T de Student (E6.1),

$$\varepsilon \sim T_k$$

  siendo k el número de grados de libertad de ε.

- Las observaciones proceden de una distribución chi cuadrado χ^2 (E6.2),

$$\varepsilon \sim \chi^2_k$$

  siendo ε el error y k el número de grados de libertad de ε.

Además, analizaremos si el estadístico sigue una distribución F de Snedecor. Para esto, veremos como si la densidad estimada del estadístico de contraste se corresponde, gráficamente, con la distribución teórica.

Por último, tendremos en cuenta como las diferencias entre los tamaños muestrales de las poblaciones pueden influir en el funcionamiento del test, aún en escenarios en los que se verifiquen todas las hipótesis del modelo.

2.4. Algoritmo

Para el cálculo de la proporción de veces que el test F rechaza H_0 se emplea el **Método de Monte Carlo**. Este es un método estadístico basado en la simulación de variables aleatorias cuando su distribución es conocida. De esta forma, se simulan en cada iteracción valores de las variables correspondientes a cada grupo, ajustando un modelo ANOVA y se calculando el nivel crítico (p-valor). Por último, se obtiene la proporción de p-valores que están por debajo de 0,05, es decir, el número de veces que rechazamos H_0. Este procedimiento se resume en el Algoritmo 1.

Algoritmo 1 Cálculo de la proporción de veces que el test F rechaza la hipótesis nula H_0.
Input: El grupo de cada individuo y el valor de la observación Y_{ij}.
Output: La proporción de veces que el test F rechaza la hipótesis nula H_0.

```
Creación de 3 grupos diferentes para la variable explicativa.
for i in 1 : 1000 do
    Simulación de los valores de Y en función del grupo al que pertenezcan.
    Ajustamos el modelo ANOVA en función de los grupos y de la variable respuesta.
    Extraemos p =p-valor del estadístico de contraste.
    if p < 0,05 then
        v[i] =TRUE
    else
        v[i] =FALSE
return mean(v)
```

2.5. Estudio del calibrado

2.5.1. Primer escenario: diseño balanceado

En el Primer Escenario, podemos apreciar que se cumplen todas las hipótesis del modelo ANOVA y que las muestras están balanceadas. Recordemos que este será el escenario de referencia, por lo que según vayamos prescindiendo de hipótesis en los diferentes modelos, cotejaremos las diferencias respecto a este.

Tras la simulación, obtenemos un $\delta = 0{,}046$, muy próximo al nivel de significación $\alpha = 0{,}05$. Por lo tanto, cuando se cumplen las hipótesis el test está bien calibrado.

Además, como podemos ver en la Figura 2.1, la densidad estimada del estadístico se acerca mucho a la densidad teórica. Por último, también podemos apreciar en la Figura 2.1 que los p-valores (niveles críticos) del estadístico siguen una distribución Uniforme $U[0, 1]$.

Podemos concluir que el test funciona correctamente en términos del calibrado cuando se cumplen las hipótesis y las muestras están balanceadas.

2.5.2. Segundo escenario: diseño desbalanceado

En el Segundo Escenario se cumplen las hipótesis del modelo, pero las muestras están desbalanceadas. Aunque esto no es una de las hipótesis del modelo, parece que puede lle-

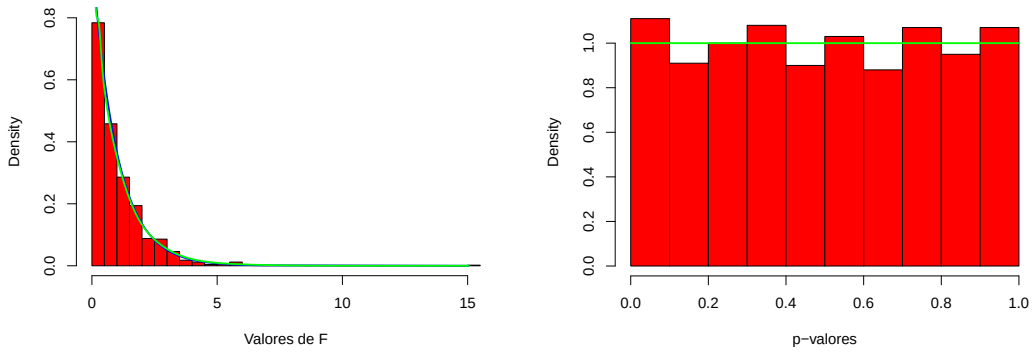

Figura 2.1: Izquierda: histograma de los valores del estadístico del Primer Escenario simulando bajo H_0. Densidad teórica(azul). Densidad estimada mediante estimación tipo núcleo (verde). Derecha: histograma de los p-valores del estadístico del Primer Escenario simulando bajo H_0. Densidad teórica(verde).

gar a afectar al calibrado del test. Por lo tanto, en este escenario, el tamaño muestral del primer grupo será $n_1 = 200$, mientras que los tamaños muestrales del segundo y del tercero tomarán la forma $n_2 = n_3 = 50$.

Simulando, obtenemos un $\delta = 0{,}053$, de nuevo próximo al nivel de significación α. De manera que, aunque las muestras estén desbalanceadas, el test está bien calibrado.

Ahora podemos apreciar en la Figura 2.2 que nuevamente la densidad estimada del estadístico está muy próxima a la teórica. Además, los p-valores del estadístico siguen teniendo una distribución muy similar a los de la Figura 2.1.

Podemos concluir que el test funciona correctamente en términos del calibrado aunque las muestras estén desbalanceadas, siempre y cuando se cumplan las hipótesis del modelo.

2.5.3. Tercer escenario: diseño con mayor variación desbalanceado

En el Tercer Escenario se siguen cumpliendo las hipótesis del modelo, pero las muestras están desbalanceadas y la varianza del error es mucho mayor. En concreto, las desviaciones típicas de los datos será $\sigma_1 = \sigma_2 = \sigma_3 = 1$, por lo que los datos tendrán una alta dispersión respecto a la media μ. Por otro lado, los tamaños muestrales serán $n_1 = 100$, $n_2 = 50$ y

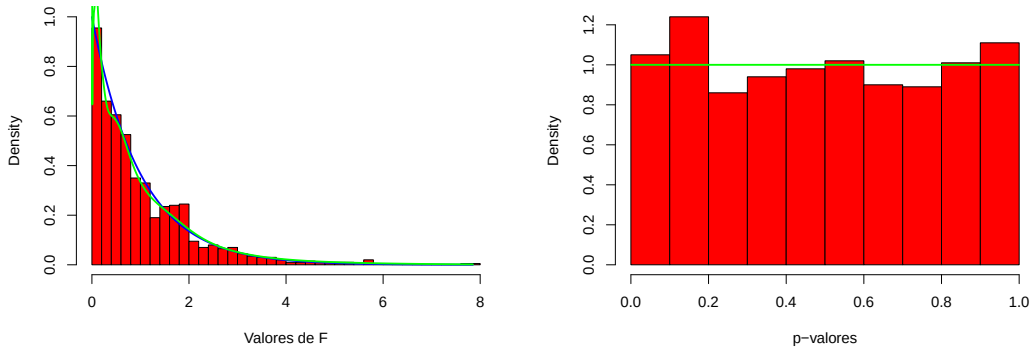

Figura 2.2: Izquierda: histograma de los valores del estadístico del Segundo Escenario simulando bajo H_0. Densidad teórica(azul). Densidad estimada mediante estimación tipo núcleo (verde). Derecha: histograma de los p-valores del estadístico del Segundo Escenario simulando bajo H_0. Densidad teórica(verde).

$n_3 = 25$.

En la simulación, obtenemos un $\delta = 0{,}052$, muy similar al nivel de significación. Por lo tanto, el test sigue estando bien calibrado.

Este escenario nos muestra que no podemos fijarnos solamente en la media de las submuestras, también hay que tener en cuenta la cantidad de datos de cada una y su desviación respecto a la media. Esto es debido a que en casos extremos en los que no tengamos demasiados datos en una muestra, y estos estén muy mucha dispersos, nos podemos encontrar con valores medios diferentes al verdadero valor medio de la población. Por último, la densidad estimada del estadístico está muy próxima a la teórica 2.3.

Podemos concluir que el test sigue funcionando correctamente en términos del calibrado aún con muestras desbalanceadas y alta variabilidad.

2.5.4. Cuarto escenario: diseño heterocedástico balanceado

Diseño heterocedástico 1

El primer apartado del Cuarto Escenario es el primero en el que no se cumple una de las hipótesis propias del modelo, en concreto veremos que ocurre cuando falla la homoce-

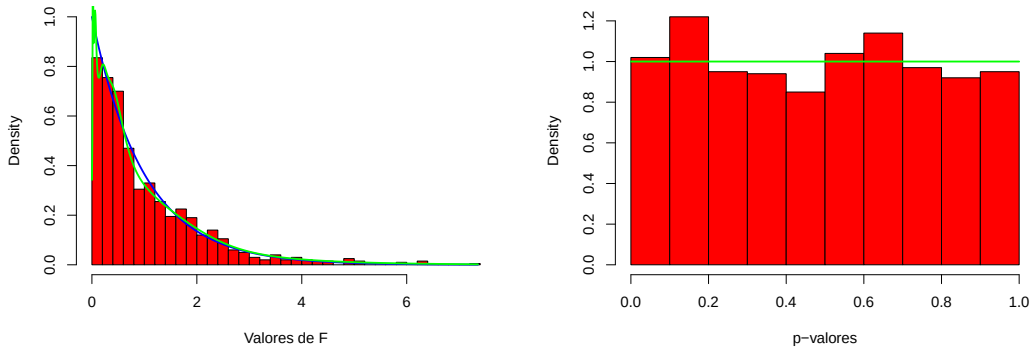

Figura 2.3: Izquierda: histograma de los valores del estadístico del Tercer Escenario simulando bajo H_0. Densidad teórica(azul). Densidad estimada mediante estimación tipo núcleo (verde). Derecha: histograma de los p-valores del estadístico del Tercer Escenario simulando bajo H_0. Densidad teórica(verde).

dasticidad. Las desviaciones típicas de los grupos 1 y 2 serán $\sigma_1 = \sigma_2 = 0{,}5$ mientras que la desviación típica del tercer grupo será $\sigma_3 = 2$. Sin embargo, los datos estarán equilibrados, teniendo en cada uno de los grupos $n = 50$ observaciones.

En la simulación, obtenemos un valor de $\delta = 0{,}074$, siendo este valor bastante diferente al nivel de significación y mayor que otros δ obtenidos. Por lo tanto, el test no está bien calibrado.

Podemos concluir diciendo que no deberíamos usar este contraste en caso de no cumplirse la hipótesis de homocedasticidad, ya que su comportamiento en términos de calibrado sería malo. Esto tendrá como consecuencia que el test rechazará la hipótesis nula una proporción de veces mayor que el nivel de significación prefijado.

Diseño heterocedástico 2

A diferencia del anterior apartado, la desviación típica del tercer grupo será menor. Concretamente, las desviaciones típicas de los grupos 1 y 2 serán $\sigma_1 = \sigma_2 = 0{,}5$ mientras que la desviación típica del tercer grupo será $\sigma_3 = 0{,}1$. Los datos seguirán estando equilibrados, teniendo en cada uno de los grupos $n = 50$ observaciones.

En la simulación, obtenemos un valor de $\delta = 0{,}064$, estando nuevamente el valor algo

lejano al nivel de significación.

Parece que en este subescenario el test tiene un comportamiento mejor en términos de calibrado que el anterior. Aún así, el test rechaza la hipótesis nula una proporción mayor de veces que el nivel de significación prefijado.

2.5.5. Quinto escenario: diseño heterocedástico desbalanceado

En el Quinto Escenario veremos qué ocurre cuando las muestras están desbalanceadas y no se cumple la hipótesis de homocedasticidad. En este escenario, las desviaciones típicas del grupo 1 y 2 serán $\sigma_1 = \sigma_2 = 0{,}5$, mientras que la del tercer grupo será $\sigma_3 = 0{,}1$. Por otra parte, el tamaño muestral será $n_1 = n_2 = 50$ y $n_3 = 200$.

Tras simular los datos, obtenemos $\delta = 0{,}278$, es decir, simulando bajo la hipótesis nula H_0, la rechazamos el 27.8 % de las veces. Como en los anteriores escenarios, el nivel de significación es $\alpha = 0{,}05$, y está muy lejos de δ. El test está claramente mal calibrado.

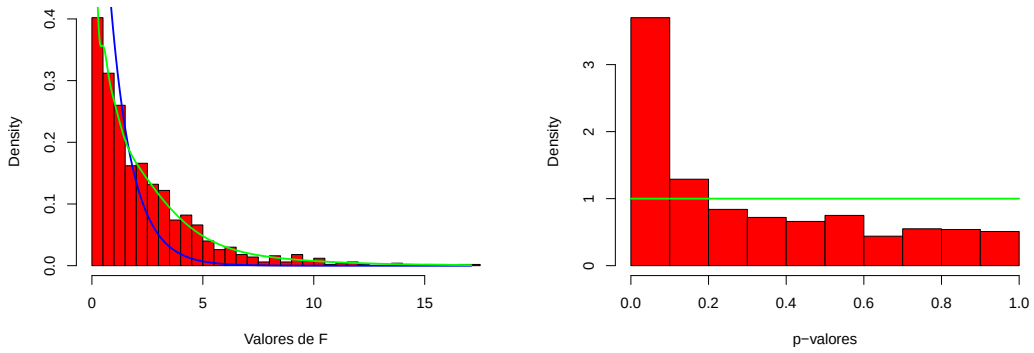

Figura 2.4: Izquierda: histograma de los valores del estadístico del Quinto Escenario. Densidad teórica(azul). Densidad estimada mediante estimación tipo núcleo (verde). Derecha: histograma de los p-valores del estadístico del Quinto escenario. Densidad teórica(verde).

En la imagen de la izquierda de la Figura 2.4 podemos apreciar diferencias muy importantes entra la función de densidad teórica y la función de densidad empírica. En la imagen de la derecha, podemos observar que hay una gran cantidad de p-valores próximos

a cero, siendo muy diferente a la distribución uniforme.

Como vimos en el escenario Diseño Desbalanceado (E2), que los datos estuviesen desbalanceados no influía en el comportamiento del test en términos de calibrado. Sin embargo, en este escenario, el test tiene un comportamiento muy diferente en términos de calibrado al desbalancear los datos, ya que aumenta considerablemente la cantidad de rechazos que se producen.

2.5.6. Sexto escenario: diseño con errores no normales balanceado

Diseño con errores T de Student

En este subescenario, observaremos qué ocurre cuando no se cumple la hipótesis de normalidad. Nuestras observaciones seguirán una distribución T de Student con 3 grados de libertad (se han cogido los mínimos posibles para que no se parezca mucho a una distribución Normal).

Al simular bajo H_0, nos encontramos con un $\delta = 0{,}051$, es decir, un valor bastante parecido al nivel de significación α. Podemos decir que el estadístico esta bien calibrado a pesar del incumplimiento de la hipótesis de normalidad.

Este es un resultado realmente llamativo, pues en este escenario logramos un δ muy próximo al nivel de significación α infringiendo directamente una de las hipótesis del modelo. De hecho, si lo comparamos con otros escenarios en los que se cumplen las hipótesis, este escenario obtiene uno de los mejores resultados en términos de calibrado.

La parte izquierda de la Figura 2.5 muestra que la densidad estimada está muy próxima a la densidad teórica, mientras que la parte derecha muestra una distribución Uniforme de los p-valores.

Podemos concluir que el test F tiene un comportamiento correcto en términos de calibrado cuando la distribución de los errores sigue siendo simétrica.

Diseño con errores asimétricos

Este será el último escenario en el que comprobaremos el calibrado de un test. En él, nos centraremos en que ocurre cuando nuestras observaciones provienen de una distribu-

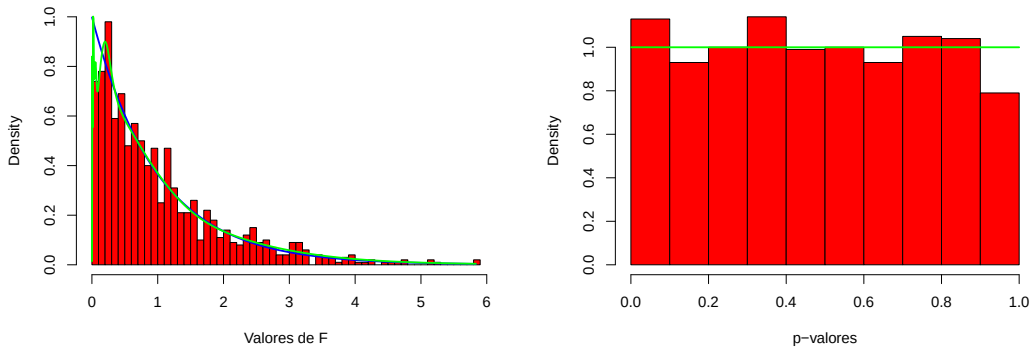

Figura 2.5: Izquierda: histograma de los valores del estadístico del Sexto Escenario simulando bajo H_0. Densidad teórica(azul). Densidad estimada mediante estimación tipo núcleo (verde). Derecha: histograma de los p-valores del estadístico del Sexto Escenario simulando bajo H_0. Densidad teórica(verde).

ción χ^2 con 5 grados de libertad.

En general, la media de una distribución χ^2 con k grados de libertad es igual a k. Debido a ello, los errores no están centrados en el cero, por lo que debemos de recentrarlos, en este caso, restando 5 unidades a las observaciones. Además, debido a la distribución que toman los errores, son asimétricos. Esto se puede apreciar con mayor claridad en la Figura 2.18.

Simulando, obtenemos un $\delta = 0{,}5$, exactamente igual a nuestro nivel de significación. Por lo tanto el estadístico está bien calibrado. Como podemos observar en la Figura 2.6, las distribuciones estimadas están muy próximas de las teóricas. Nuevamente, parece que pese al quebrantamiento de la hipótesis de normalidad, el test funciona perfectamente en términos de calibrado.

Podemos concluir que el estadístico tiene un comportamiento correcto en términos de calibrado.

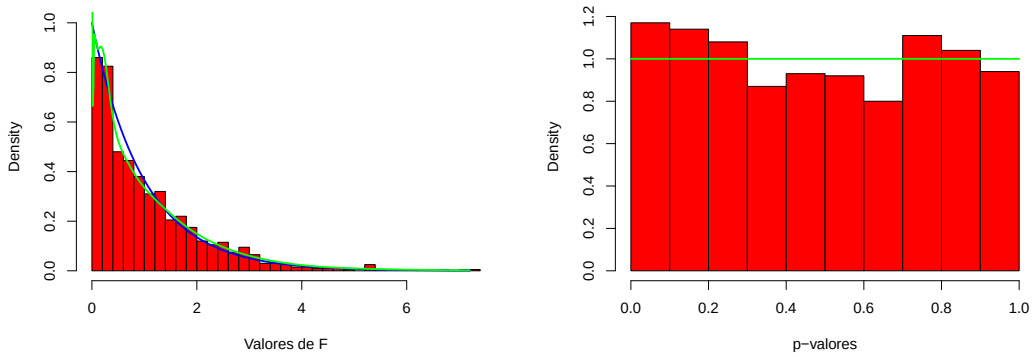

Figura 2.6: Izquierda: histograma de los valores del estadístico del Séptimo Escenario simulando bajo H_0. Densidad teórica(azul). Densidad estimada mediante estimación tipo núcleo (verde). Derecha: histograma de los p-valores del estadístico del Séptimo Escenario simulando bajo H_0. Densidad teórica(verde).

2.6. Estudio de la potencia

En la Sección 2.3 mencionamos que la potencia del test variará dependiendo de diferentes factores, como el **número de observaciones** o la **diferencia de medias**. En consecuencia, consideraremos diferentes tamaños muestrales y diferentes medias, y observaremos como varía la potencia del test en función de estas.

Por ello, las medias de los grupos 1 y 2 serán $\mu_1 = \mu_2 = 1$, mientras que iremos variando la media del tercer grupo. De igual forma, una vez fijada la media del tercer grupo, aumentaremos el número de observaciones en cada grupo. Una vez realizado esto, veremos como va variando la potencia del test F.

2.6.1. Primer escenario: diseño balanceado

Como podemos observar en la imagen derecha de la Figura 2.8, a medida que aumenta la diferencia de medias entre el tercer grupo respecto al primero y al segundo, más rápido se aproxima la potencia del test a 1. Además, también podemos apreciar que el test se comporta de forma adecuada en términos de potencia, ya que esta se incrementa según aumenta el tamaño muestral.

Así, si tomamos un tamaño muestral de $n_1 = n_2 = n_3 = 50$ y $\mu_3 = 1{,}3$, obtenemos $\theta = 0{,}889$, es decir, el test rechaza la igualdad de medias un 88.9 % de las veces, obteniendo el test una potencia alta.

Por el contrario, si tomásemos un tamaño muestral de $n_1 = n_2 = n_3 = 50$, pero $\mu_3 = 1{,}1$, obtendríamos un $\theta = 0{,}159$, una potencia inferior a la anterior, como cabe esperar, dado que el valor de la media del tercer grupo está más próximo.

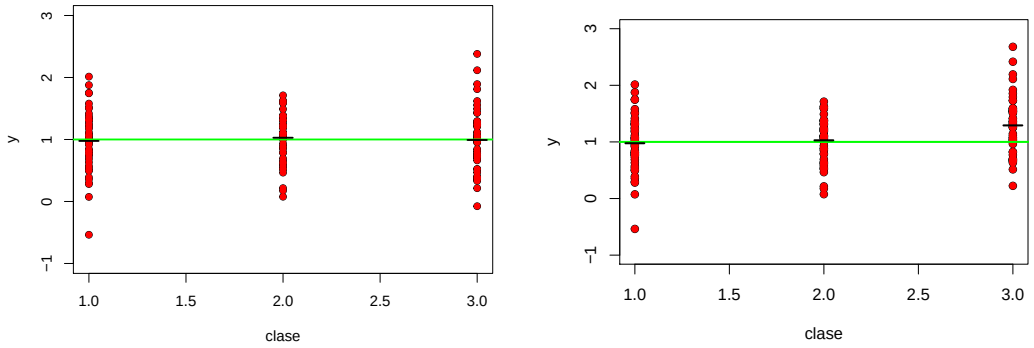

Figura 2.7: Diagrama de dispersión de la variable respuesta Y en función de la clase a la que pertenezcan, simulado mediante técnicas de Montecarlo. Izquierda: datos simulados bajo H_0 en el Primer Escenario. Derecha: datos simulados bajo H_a en el Primer Escenario. La línea verde representa la media teórica, mientras que las líneas negras de cada clase representan la media muestral de dicha clase.

Para finalizar con este escenario, en la Figura 2.7 podemos apreciar como los datos de la tercera clase son mayores al simular bajo H_a, debido a esto, el test F detecta diferencias entre las medias en la clase 3 con respecto a las otras dos.

2.6.2. Segundo escenario: diseño desbalanceado

De nuevo simularemos bajo la hipótesis alternativa H_a, pero a diferencia del anterior escenario, las muestras estarán desbalanceadas. Para ello tomamos un tamaño muestral $n_1 = n$ en el primer grupo, mientras que el tamaño de los grupos 2 y 3 será $n_2 = n_3 = n/4$.

Podemos observar en la imagen izquierda de la Figura 2.8 [1]como aumenta la potencia

[1]Con el fin de comparar el comportamiento en términos de potencia entre el Primer Escenario: diseño

del test según aumenta el tamaño muestral, lo que vuelve a indicar un buen comportamiento en términos de potencia.

Sin embargo, podemos apreciar una diferencia notable, y es que en el diseño desbalanceado necesitamos un tamaño muestral mayor para llegar a una potencia próxima a 1. Así, en el Primer Escenario (E1), cuando $n = 1000$ (3000 datos totales), la potencia del test está próxima a 1. De igual forma, en el Segundo Escenario (E2), cuando tomamos el valor $n = 2000$ (3000 datos totales), la potencia del test toma un valor algo más pequeño .Esto muestra la importancia de que nuestros datos estén balanceados, pudiendo llegar a falsas conclusiones al realizar un estudio poblacional.

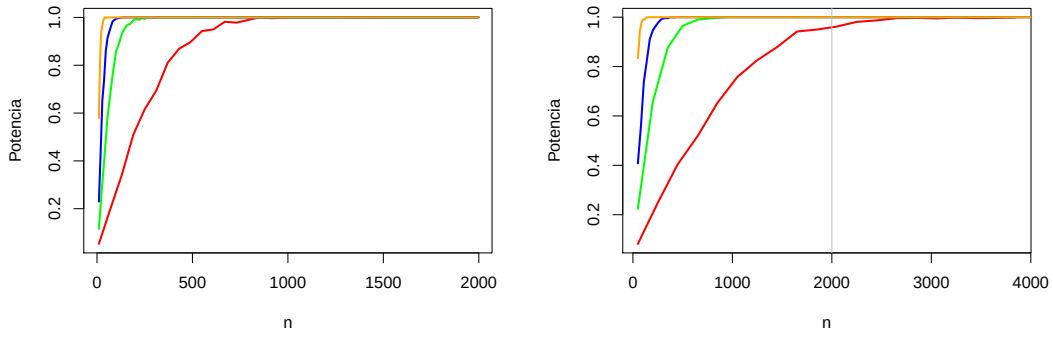

Figura 2.8: Gráficas de la potencia del test del Primer Escenario (gráfica izquierda) y del Segundo Escenario (gráfica derecha) en función del tamaño muestral. Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

2.6.3. Tercer escenario: diseño desbalanceado con mayor variación

En el estudio de la potencia del Tercer Escenario se pone de manifiesto la importancia de la varianza del error. Simularemos bajo H_a tomando diferentes tamaños muestrales, donde $n_1 = n$, $n_2 = n/2$ y $n_3 = n/4$. En este escenario la desviación típica será mayor,

balanceado con los demás escenarios, se ha marcado el valor $n = 2000$ con una línea gris. Esto permite obtener una perspectiva del tamaño muestral en los escenarios en los que se necesita un mayor número de datos. Esta información se representa en todas las gráficas de potencia a los largo de la memoria.

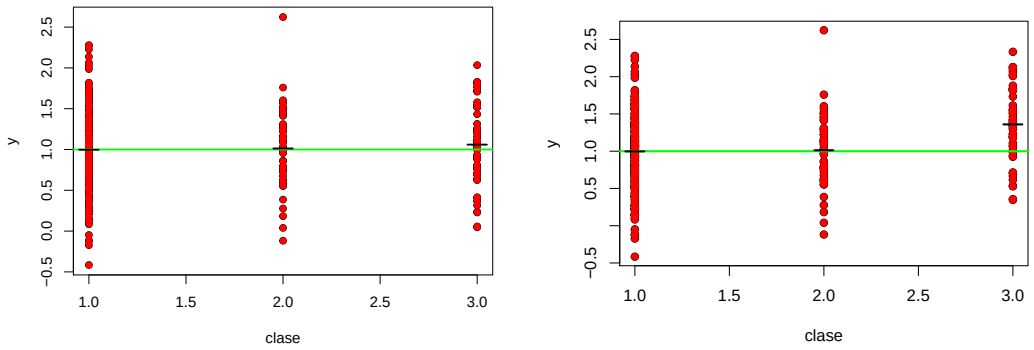

Figura 2.9: Diagrama de dispersión de la variable respuesta Y en función de la clase a la que pertenezcan, simulado mediante técnicas de Montecarlo. Izquierda: datos simulados bajo H_0 en el Segundo Escenario. Derecha: datos simulados bajo H_a en el Segundo Escenario. La línea verde representa la media teórica, mientras que las líneas negras de cada clase representan la media muestral de dicha clase.

tomando un valor de $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

En la Figura 2.10, podemos comparar el comportamiento en términos de potencia del Primer Escenario con respecto al Tercer Escenario, ahí se puede apreciar que en el Tercer Escenario (E3) necesitamos un tamaño muestral mayor para obtener una potencia próxima a 1. Además, como podemos ver en la imagen derecha de la Figura 2.10, el Tercer Escenario muestra un buen comportamiento en términos de potencia, ya que esta se acerca a 1 según aumenta el tamaño muestral.

Parece que una mayor dispersión en los datos provoca que se pierda potencia en el test, por lo que necesitaremos un tamaño muestral mayor para que el test tenga una potencia elevada.

2.6.4. Cuarto escenario: diseño heterocedástico balanceado

Diseño heterocedástico 1

Ahora llevaremos a cabo el estudio de la potencia del Cuarto Escenario. En este caso, los tamaños muestrales serán iguales, tomando los valores $n_1 = n_2 = n_3 = n$. Además, no se cumple la hipótesis de homocedasticidad del modelo, siendo las desviaciones típicas de

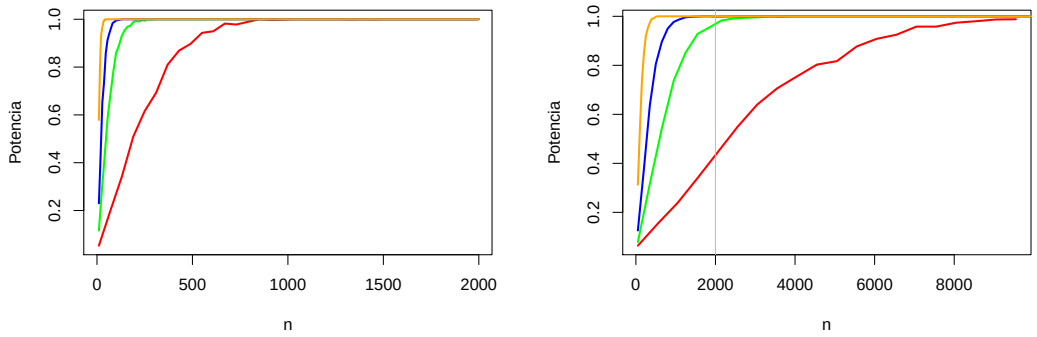

Figura 2.10: Gráficas de la potencia del test del Primer Escenario (gráfica izquierda) y del Tercer Escenario (gráfica derecha) en función del tamaño muestral. Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

los grupos $\sigma_1 = \sigma_2 = 0{,}5$ y $\sigma_3 = 2$.

Compararemos este subescenario nuevamente con el Primer Escenario (E1) con el objetivo de ver como afecta una mayor dispersión en uno de los grupos, pues la única diferencia entre ellos es una mayor dispersión en el tercer grupo. Como podemos ver en la imagen derecha de la Figura 2.12, en el Cuarto Escenario: diseño heterocedástico 1 (E4.1), son necesarios muchos más datos para que la potencia esté próxima a 1. De esta forma, tomando un $n = 2000$ en el Primer Escenario (E1) tenemos una potencia próxima a 1, mientras que al tomar el mismo n en el Cuarto Escenario: diseño heterocedástico 1 (E4.1), la potencia es menor que 0,5.

Este resultado, parece mostrar que cuando la dispersión de uno de los grupos es mayor, el test obtiene peores resultados en términos de potencia, por lo que antes de realizar el test F sería conveniente validar la hipótesis de homocedasticidad usando el test de Levene.

Diseño heterocedástico 2

Haremos un estudio de simulación similar al del anterior escenario, pero en este caso la desviación típica del tercer grupo será menor. Por lo tanto, los tamaños muestrales serán iguales, tomando los valores $n_1 = n_2 = n_3 = n$. Además, no se cumple la hipótesis de

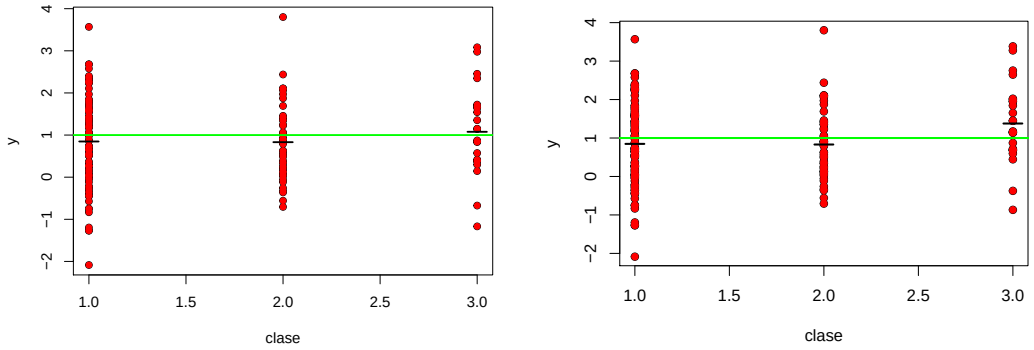

Figura 2.11: Diagrama de dispersión de la variable respuesta Y en función de la clase a la que pertenezcan, simulado mediante técnicas de Montecarlo. Izquierda: datos simulados bajo H_0 en el Tercer Escenario. Derecha: datos simulados bajo H_a en el Tercer Escenario. La línea verde representa la media teórica, mientras que las líneas negras de cada clase representan la media muestral de dicha clase.

homocedasticidad del modelo, siendo las desviaciones típicas de los grupos $\sigma_1 = \sigma_2 = 0{,}5$ y $\sigma_3 = 0{,}1$.

Nuevamente compararemos este subescenario con el Primer Escenario (E1), ahora con el objetivo de ver como afectaría una menor dispersión en uno de los grupos, ya que la única diferencia entre ellos es una menor dispersión en el tercer grupo. En la imagen izquierda de la Figura 2.12 podemos apreciar que ahora el test presenta un mejor comportamiento en términos de potencia, ya que se aproxima antes a 1. De hecho, el test presenta en este escenario un mejor resultado en términos de potencia que el escenario Diseño balanceado (E1), que es el que tomamos como referencia y cumple todas las hipótesis.

Esto es llamativo, pues es el escenario con mejor comportamiento en términos de potencia. Probablemente se deba a que en estas condiciones, el test tiende a rechazar H_0.

2.6.5. Quinto escenario: diseño heterocedástico desbalanceado

En este escenario, las condiciones serán ligeramente diferentes al anterior. Con el objetivo de ver como afecta, en términos de la potencia, que las muestras estén desbalanceadas a los datos, mantendremos los valores de las desviaciones típicas, mientras que reduciremos el tamaño muestral del primer y segundo grupo, tomando estos los valores $n_1 = n_2 = n/4$

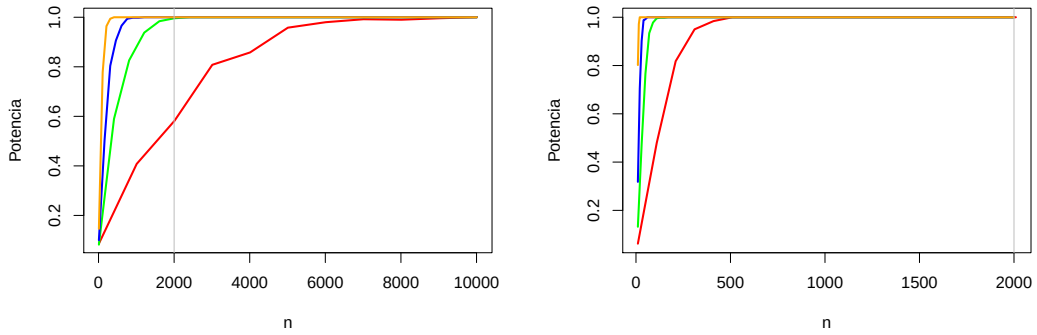

Figura 2.12: Gráficas de la potencia del test del Cuarto Escenario: diseño heterocedástico 1 (gráfica izquierda) y del Cuarto Escenario: diseño heterocedástico 2 (gráfica derecha) en función del tamaño muestral. Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

y $n_2 = n$.

Debido a esto, será conveniente comparar este escenario con el Cuarto Escenario: diseño heterocedástico (E4.2). Esto se puede ver reflejado en la Figura 2.13, donde en el escenario (E4.2), con $n = 500$ (tenemos 1500 datos totales), alcanzamos valores de potencia próximos a 1. De igual forma, en este escenario, alcanzamos dichos valores con $n = 1000$ (de nuevo 1500 datos totales). En este caso parece que el desbalanceado de las muestras no afecta demasiado a la potencia del test.

Por otra parte, como podemos ver en la imagen izquierda de la Figura 2.13, el test sigue manteniendo un buen comportamiento en términos de potencia, aproximándose a 1 a medida que aumenta el tamaño muestral.

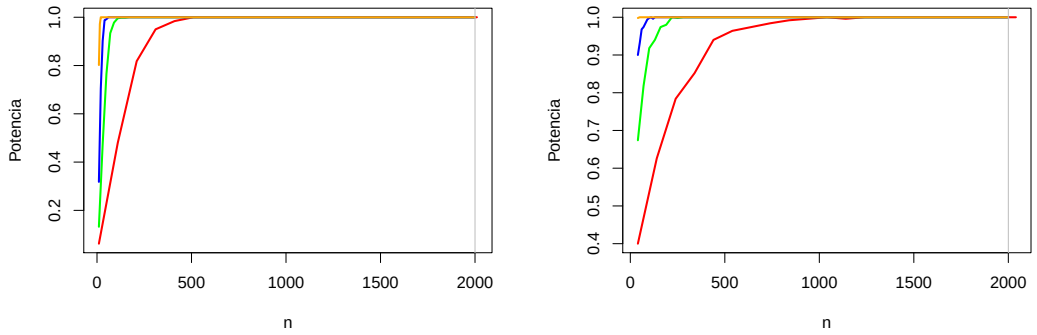

Figura 2.13: Gráficas de la potencia del test del Cuarto Escenario: diseño heterocedástico 2 (gráfica izquierda) y del Quinto Escenario (gráfica derecha) en función del tamaño muestral. Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

2.6.6. Sexto escenario: diseño con errores no normales balanceado

Diseño con errores T de Student

En los dos últimos subescenarios ponemos a prueba la hipótesis de normalidad. En este, los errores siguen una distribución T de Student, lo cual da lugar a errores simétricos pero no normales.

En la Figura 2.15 podemos observar que el subescenario (E6.1) muestra un buen comportamiento en términos de potencia, pero necesitamos un mayor tamaño muestral que en el escenario diseño balanceado (E1) para alcanzar una potencia próxima a 1.

Esto parece indicar que cuando los errores siguen una distribución T de Student en vez de una distribución normal el test obtiene peores resultados en términos de potencia. Debido a esto, sería conveniente validar la hipótesis de normalidad en cada grupo usando el test de Shapiro-Wilks o el de Lilliefors.

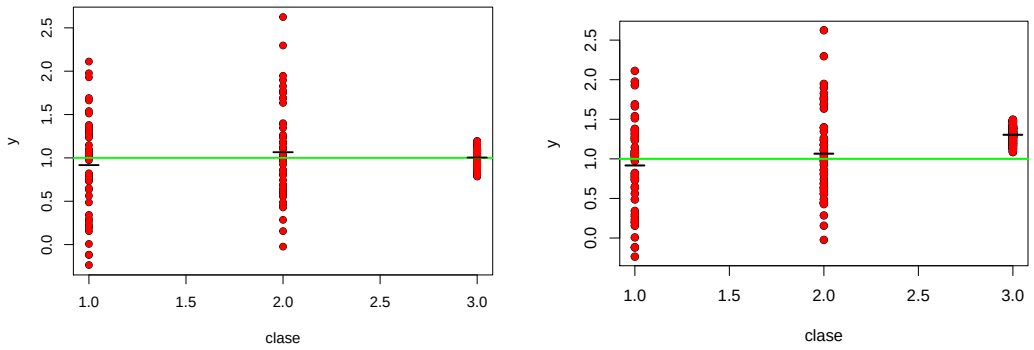

Figura 2.14: Diagrama de dispersión de la variable respuesta Y en función de la clase a la que pertenezcan, simulado mediante técnicas de Montecarlo. Izquierda: datos simulados bajo H_0 en el Quinto Escenario. Derecha: datos simulados bajo H_a en el Quinto Escenario. La línea verde representa la media teórica, mientras que las líneas negras de cada clase representan la media muestral de dicha clase.

Diseño con errores asimétricos

Por último, tomaremos errores que siguen una distribución χ^2, con 5 grados de libertad. Como dijimos en el estudio del calibrado de este escenario, debemos de recentrar los errores para que su media sea igual a 0. Además, la distribución de los errores es asimétrica.

En la imagen izquierda de la Figura 2.17, podemos observar cómo es el escenario en el que se necesita un mayor tamaño muestral para alcanzar una potencia próxima a 1. De esta forma, si $\mu_3 = 1{,}5$ y $n = 1000$ en las condiciones del Primer escenario, la potencia del test estaría próxima a 1, lo que implica que seríamos capaces de rechazar H_0 en la mayoría de las situaciones. Los mismos datos en las condiciones del Sexto Escenario: diseño con errores asimétricos (E6.2), tendrían asociados una potencia menor de $\theta = 0{,}2$, por lo que no rechazaríamos H_0 en la mayoría de los casos. Esto índica que no deberíamos usar datos no normales, ya que en muchas situaciones no seremos capaces de rechazar H_0.

Además, en la Figura 2.18 podemos apreciar la asimetría de los datos, así como el ligero efecto que tiene en las gráficas la diferencia de medias. Podemos concluir que para el adecuado comportamiento del test, este necesitaría un tamaño muestral muy grande.

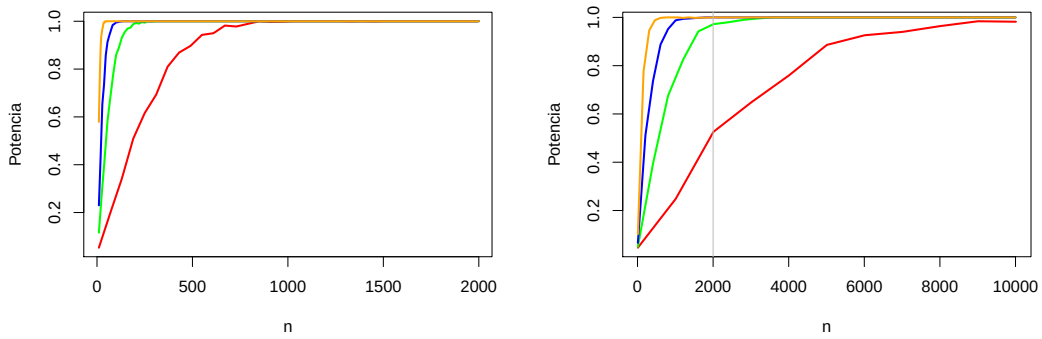

Figura 2.15: Gráficas de la potencia del test del Primer Escenario (gráfica izquierda) y del Sexto Escenario (gráfica derecha) en función del tamaño muestral. Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

2.7. Conclusiones

Tras realizar un estudio exhaustivo de simulación, podemos observar lo siguiente:

- **Datos desbalanceados**: es recomendable usar muestras balanceadas. En caso de que la muestra esté desbalanceada y los datos cumplan con las hipótesis del modelo ANOVA, el test F obtiene buenos resultados en términos de calibrado y potencia. Por otro lado, si la muestra está desbalanceada pero no cumple alguna de las hipótesis del modelo, el test empeora sus resultados notablemente.

- **Datos heterocedásticos**: podemos apreciar que:

  - Si el diseño es equilibrado, el modelo ANOVA es bastante robusto (esto lo podemos ver en el Cuarto Escenario) ya que la proporción de rechazos δ tiene tendencia a ser mayor que α, pero sin alejarse mucho. En lo que se refiere al comportamiento del test en términos de potencia, parece que si la varianza de uno de los grupos es mayor, el test obtiene peores resultados, mientras que si por el contrario la varianza de uno de los grupos es menor, obtiene resultados muy similares en términos de potencia al escenario diseño balanceado.
  - En diseños no equilibrados, la falta de homocedasticidad tiene mayor impacto. Si los grupos de menor tamaño son los que presentan una mayor desviación típica, habrá una mayor cantidad de falsos positivos (esto es lo que ocurre en

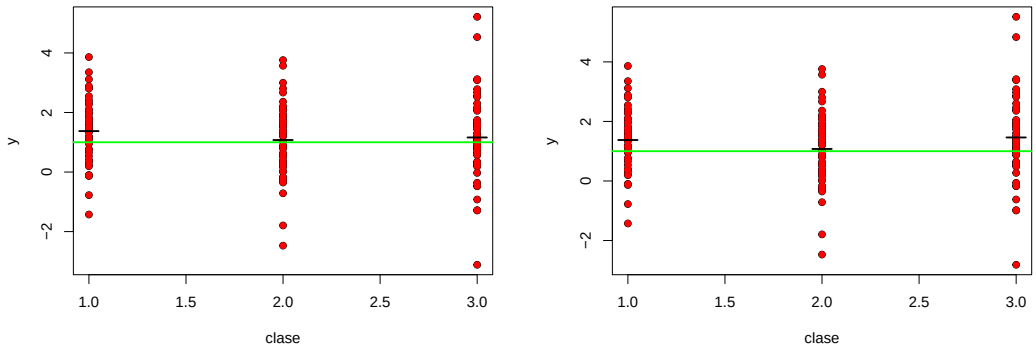

Figura 2.16: Diagrama de dispersión de la variable respuesta Y en función de la clase a la que pertenezcan, simulado mediante técnicas de Monte Carlo. Izquierda: datos simulados bajo H_0 en el Sexto Escenario. Derecha: datos simulados bajo H_a en el Sexto Escenario. La línea verde representa la media teórica, mientras que las líneas negras de cada clase representan la media muestral de dicha clase.

el Quinto Escenario). Si por el contrario, los grupos de mayor tamaño tienen mayor desviación típica, aumentaran los falsos negativos. Además, el test obtiene resultados muy pobres en términos de calibrado.

Sería conveniente realizar una validación de la hipótesis de homocedasticidad para ver cuál es el escenario al que nos enfrentamos. Esto lo podríamos comprobar realizando el test de Levene, el cual sirve para comprobar si la varianza de los grupos es la misma. Además, también deberíamos de tener en cuenta si nuestros datos están balanceados.

- **Datos no normales**: el modelo ANOVA funciona bien en términos de calibrado cuando los errores no siguen una distribución Normal dentro de cada grupo. Sin embargo, el test tiene una baja potencia. Sería adecuado realizar una correcta validación de la hipótesis de Normalidad para ver en que contexto nos encontramos. Esto lo podríamos llevar a cabo realizando el test de Shapiro-Wilks o el de Lilliefors para contrastar la normalidad.

- **Tamaño muestral**: cuanto mayor sea el tamaño muestral, más fiables serán nuestros resultados. Esto se evidencia en el estudio de la potencia del Primer Escenario, en el cual podemos apreciar que según aumenta n, aumenta la potencia del test.

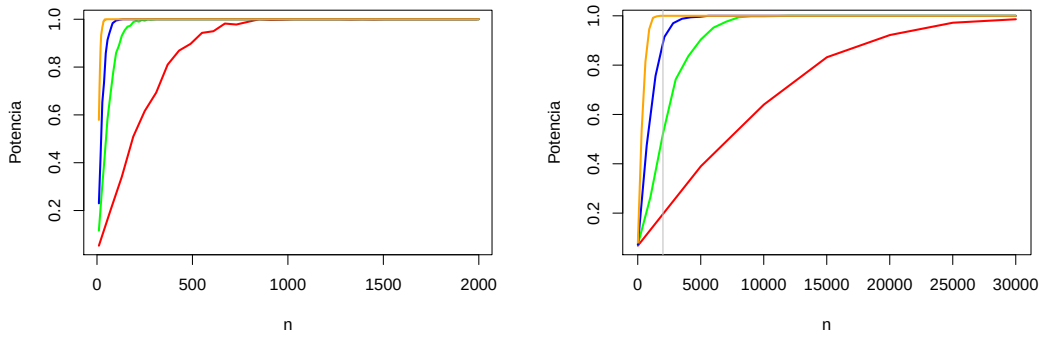

Figura 2.17: Gráficas de la potencia del test del Primer Escenario (gráfica izquierda) y del Séptimo Escenario (gráfica derecha) en función del tamaño muestral. Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

2.8. Una alternativa no paramétrica al test ANOVA

Tras realizar el estudio de simulación, podemos observar que el test F presenta problemas en términos de potencia y calibrado cuando no se cumplen las hipótesis de homocedasticidad y de normalidad. Estos problemas se ven agravados cuando las muestras están desbalanceadas. Como comentamos en la Sección 2.7, sería apropiado realizar una conveniente validación de las hipótesis de homocedasticidad y de normalidad antes de aplicar el test. En esta sección presentaremos brevemente una alternativa al modelo ANOVA cuando no se cumplen la hipótesis de normalidad.

El Test Kruskal-Wallis (Corder y Foreman (2011)), también conocido como test H, es un método no paramétrico alternativo al test F, el cual emplea rangos para contrastar la hipótesis de que k muestras han sido obtenidas de la misma población. A diferencia de este, el test de Kruskal-Wallis no asume normalidad en los datos, pudiendo formular el contraste de hipótesis del siguiente modo:

$$H_0 : \text{Las muestras provienen de una misma población.}$$

$$H_a : \text{Al menos una muestra proviene de una población diferente.}$$

Supongamos que disponemos de k grupos con uno de ellos con n observaciones. Entonces

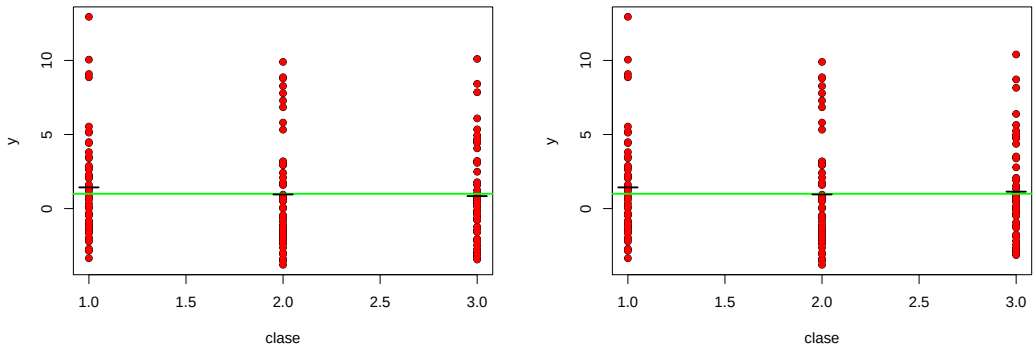

Figura 2.18: Diagrama de dispersión de la variable respuesta Y en función de la clase a la que pertenezcan, simulado mediante técnicas de Montecarlo. Izquierda: datos simulados bajo H_0 en el Séptimo Escenario. Derecha: datos simulados bajo H_a en el Séptimo Escenario. La línea verde representa la media teórica, mientras que las líneas negras de cada clase representan la media muestral de dicha clase.

el estadístico H se calcularía como:

$$H = (N-1)\frac{\sum_{i=1}^{k} n_i(\bar{r}_{i\bullet} - \bar{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \tag{2.1}$$

donde:

- n_i es el numero de observaciones en el grupo i.
- r_{ij} es igual al rango de la observación j en el grupo i.
- N representa el número total de observaciones entre todos los grupos.
- $\bar{r}_{i\bullet}$ es el rango medio del grupo i.
- $\bar{r}$ es el promedio de r_{ij}.

Al igual que el modelo ANOVA, seguimos considerando observaciones independientes de una variable continua que se observa en varios grupos. Además, dado que en la hipótesis nula estamos asumiendo que los grupos pertenecen a una misma población, se sigue manteniendo la hipótesis de homocedasticidad. Por último, bajo H_0, se asume que todos los grupos siguen una misma distribución.

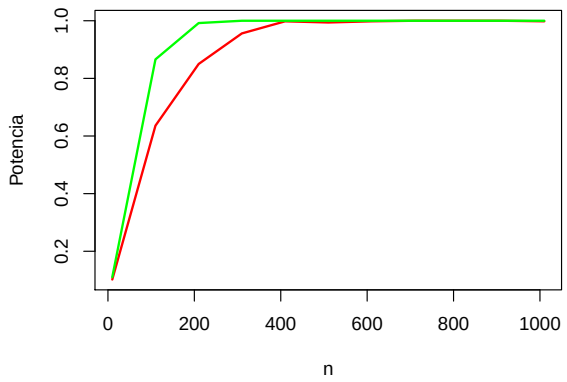

Figura 2.19: Diagrama de dispersión de la potencia en función del tamaño muestral cuando los errores siguen una distribución T de Student y las muestras están pareadas. Línea roja: potencia del test ANOVA. Línea verde: potencia del test Kruskal-Wallis.

En la Figura 2.19, podemos observar la diferencia entre la potencia de el test F y el test Kruskal-Wallis cuando toman el mismo tamaño muestral y los errores siguen una distribución T de Student. Es evidente que el test Kruskal-Wallis obtiene mejores valores de potencia cuando el tamaño muestral es el mismo.

Capítulo 3

Estudios de simulación sobre el modelo ANCOVA

En este capítulo pondremos a prueba el test F en dos estadísticos de contraste diferentes en el contexto del modelo ANCOVA. Para ello realizaremos los test de no efecto. En ambos contrastes iremos suprimiendo distintas hipótesis del modelo y viendo como responden los test.

Para llevar esto a cabo, en la Sección 3.1 introduciremos los diferentes contrastes a tener en cuenta. En la Sección 3.2 presentaremos los diferentes escenarios de simulación que testaremos. Por su parte, la Sección 3.3 estará dedicada a narrar la metodología utilizada para realizar el estudio de simulación. Más tarde, en las Secciones 3.5.1 - 3.5.2 presentaremos e interpretaremos los resultados del estudio de simulación en términos del calibrado y de la potencia del contraste de la variable continua y de la variable discreta. Por último, en la Sección 3.6, extraeremos las conclusiones de los estudios realizado.

3.1. Contrastes de efecto

En el Capítulo 2, usamos el test F para contrastar si el papel del grupo tiene algún efecto en Y. En el modelo ANCOVA (Maxwell et al. (2017)), además de una variable explicativa discreta, tenemos una variable explicativa continua, por lo tanto, tendremos que contrastar el papel de ambas variables.

Recordemos que el contraste del efecto de la variable continua era:

$$H_0 : \gamma = 0$$

$$H_a : \gamma \neq 0$$

mientras que el de la variable discreta toma la forma:

$$H_0 : \eta_i = 0 \ \forall i$$

$$H_a : \exists i \text{ tal que } \eta_i \neq 0$$

donde γ y η_i son los especificados en el modelo (1.11).

Análogamente al Capítulo 2, fijado un nivel de significación $\alpha \in (0, 1)$, evaluaremos el comportamiento del test en términos de **calibrado** y **potencia**, viendo el impacto de la ausencia de alguna de las hipótesis del modelo ANCOVA.

3.2. Escenarios de simulación

Empezaremos evaluando los test en términos de calibrado y potencia. Presentaremos 4 escenarios de simulación diferentes, donde se han considerado $I = 3$ clases:

1. **Diseño homocedástico balanceado (A1)**: las desviaciones típicas de los grupos 1, 2 y 3 serán $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$ respectivamente. El tamaño de muestra en el estudio del calibrado será igual para los 3 grupos, siendo este $n_1 = n_2 = n_3 = 50$.

2. **Diseño homocedástico desbalanceado (A2)**: las desviaciones típicas de los grupos 1, 2 y 3 serán $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$ respectivamente. El tamaño de muestra en el estudio del calibrado será diferente en este caso, siendo este $n_1 = n_2 = 10$, mientras que $n_3 = 80$.

3. **Diseño heterocedástico balanceado (A3)**: las desviaciones típicas de los grupos en este caso serán diferentes, siendo $\sigma_1 = \sigma_2 = 0{,}5$ la desviación típica de los grupos 1 y 2, así como $\sigma_3 = 2$ la del grupo 3. En este escenario, en el estudio del calibrado, las muestras estarán balanceadas, tomando las tres el valor $n_1 = n_2 = n_3 = 50$. Es importante mencionar que será el primer escenario en el que no se cumple una de las hipótesis del modelo ANCOVA.

4. **Diseño heterocedástico desbalanceado (A4)**: las desviaciones típicas de los grupos en este caso serán diferentes, siendo $\sigma_1 = \sigma_2 = 0{,}5$ la desviación típica de los grupos 1 y 2, así como $\sigma_3 = 2$ la del grupo 3. En este escenario, en el estudio del calibrado, las muestras estarán desbalanceadas, tomando los valores $n_1 = n_2 = 10$ y $n_3 = 80$.

Como hemos mencionado, evaluaremos el test en términos de calibrado y potencia, por lo que debemos de simular bajo H_0 y bajo H_a. Como tenemos dos contrastes de hipótesis diferentes, primero realizaremos el contraste de la variable continua y luego el de la discreta. En cada uno de ellos, dejaremos fijo el valor de los parámetros de la variable que no estemos contrastando y evaluaremos la otra. Esto lo haremos estudiando primero el comportamiento en términos de calibrado, y posteriormente, tomando diferentes valores de la variable a contrastar y diferentes tamaños muestrales.

3.3. Código y metodología

Como en la Sección 2.3, se ha hecho uso del lenguaje de programación R (R Core Team (2019)). Hemos empleado un método similar al del Capítulo 2, ya que se ha simulado $M = 1000$ veces cada uno de los escenarios mediante **técnicas de Montecarlo**.

Como en este caso tenemos que testear dos contrastes, comprobaremos el buen funcionamiento en términos de calibrado de ambos, fijando $\alpha = 0{,}05$ en ambos test y observando cual es la probabilidad de que el test rechace las hipótesis nulas H_0. A la probabilidad de que el test rechace el contraste de la variable continua o el de la variable discreta cuando simulamos bajo H_0 lo denotaremos con la letra δ[1].

Para el estudio de la potencia debemos de simular bajo H_a. De esta forma, llamaremos θ a la probabilidad de que el test rechace la hipótesis nula en el contraste del efecto de la variable continua o en el de la discreta, condicionado a que estamos simulando bajo H_a .

En este caso, como tenemos que estudiar el comportamiento de dos test diferentes, haremos lo siguiente:

- **Contraste de no efecto de la variable continua**: evaluaremos el efecto de la variable continua fijando el valor de los parámetros de la variable discreta η_i. Para esto, primero analizaremos el comportamiento del test en términos de calibrado, para, a posteriori, testear el comportamiento del test en términos de potencia, tomando diferentes valores de la pendiente γ y del tamaño muestral n.
- **Contraste de no efecto de la variable discreta**: evaluaremos el efecto de la variable discreta fijando en este caso el valor de los parámetros de la variable continua

[1]En este caso usaremos la notación δ para referirnos a cuando estamos simulando bajo H_0 y la notación θ para referirnos a cuando estamos simulando bajo H_a. Por lo demás, representan lo mismo, la probabilidad de rechazar la hipótesis nula H_0.

γ. Para hacer esto, analizaremos el comportamiento del test en términos de calibrado, y posteriormente, observaremos el comportamiento del test en términos de potencia, tomando diferentes valores de las medias de los grupos $\mu 1$, μ_2 y μ_3, además de variar el tamaño muestral n.

3.4. Algoritmo

Para el cálculo de la proporción de veces que el test F rechaza H_0 se emplea nuevamente el **Método de Monte Carlo**. Realizaremos un proceso similar al del Capítulo 2, pero ahora, además de contrastar la variable discreta, también debemos de contrastar la variable continua. De esta forma, se simulan en cada iteracción valores de las variables correspondientes a cada grupo y una variable continua, ajustando un modelo ANCOVA y calculando el nivel crítico de ambos contrastes (p-valor). Por último, se obtiene la proporción de p-valores que están por debajo de 0,05, es decir, el número de veces que rechazamos H_0. Este procedimiento se resume en el Algoritmo 2.

3.5. Test de no efecto

Como hemos comentado, tenemos que contrastar el efecto de la variable continua y el de la variable discreta en dos test diferentes. Por ello, fijaremos el efecto de una de las variables y observaremos como evoluciona la otra en términos de calibrado y potencia al tomar distintos valores.

3.5.1. Contraste de no efecto de la variable continua

En el contrate de no efecto de la variable continua fijaremos los valores de las medias de los tres grupos $\mu_1 = \mu_2 = \mu_3 = 1$ [2] y tomaremos diferentes valores de la pendiente de la recta de regresión γ.

Primer escenario: diseño homocedástico balanceado

En este escenario las desviaciones típicas de los grupos serán $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$. Además, es importante mencionar que se cumplen todas las hipótesis del modelo ANCOVA.

[2] De ahora en adelante, consideraremos $\mu_i = \mu + \eta_i$, donde μ es una constante y η_i representa al efecto del grupo i.

Algoritmo 2 Cálculo de la proporción de veces que el test F rechaza las hipótesis nulas H_0.

Input: El grupo de cada individuo, una variable continua y el valor de la observación Y_{ij}.

Output: La proporción de veces que el test F rechaza las hipótesis nulas H_0.

Creación de 3 grupos diferentes para la variable explicativa.
for i in $1:1000$ **do**
    Simulación de los valores de Y en función del grupo al que pertenezcan.
    Simulación de los valores de X.
    Ajustamos el modelo ANCOVA en función de los grupos, de la variable continua y de la variable respuesta.
    Extraemos p_1 =p-valor del estadístico del contraste de la variable continua.
    if $p_1 < 0{,}05$ **then**
        $v_1[i]$ =TRUE
    else
        $v_1[i]$ =FALSE
    Extraemos p_2 =p-valor del estadístico del contraste de la variable discreta.
    if $p_2 < 0{,}05$ **then**
        $v_2[i]$ =TRUE
    else
        $v_2[i]$ =FALSE
return $mean(v_1)$, $mean(v_2)$

1. $\boldsymbol{\gamma = 0}$: tras simular obtenemos el valor $\delta = 0{,}046$, cercano al nivel de significación $\alpha = 0{,}05$. Podemos deducir de esto que el test está bien calibrado.

2. $\boldsymbol{\gamma \neq 0}$: en este caso, estudiaremos diferentes valores de γ y veremos como se comporta la potencia del test al aumentar el propio γ y el tamaño muestral n.

En la Figura 3.1 podemos observar que cuanto mayor es la pendiente de la recta de regresión, necesitamos un menor tamaño muestral para tener una mayor potencia en el test y así identificar que existe un efecto de la variable continua. Así, si $\gamma = 0{,}05$, necesitamos aproximadamente un tamaño muestral $n_1 = n_2 = n_3 = 500$ para tener una potencia cercana a 1. Sin embargo, si aumentamos la pendiente de la recta de regresión hasta $\gamma = 0{,}2$, con un n cercano a 50 obtendríamos una potencia igual a 1.

Por lo tanto, podemos deducir que en el contraste de la variable continua, cuanto mayor sea el valor absoluto de la pendiente de la recta de regresión γ, mayor será la potencia del test. De igual forma, cuanto mayor sea el tamaño muestral, más potencia tendrá el test F. Además, podemos concluir que el test tiene un buen comportamiento en términos de potencia, ya que al aumentar el tamaño muestral, la potencia del test se aproxima a 1.

Segundo escenario: diseño homocedástico desbalanceado

En este escenario se mantendrán las desviaciones típicas de los grupos, tomando el valor $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$, pero a diferencia del Primer Escenario las muestras estarán desbalanceadas.

1. $\boldsymbol{\gamma = 0}$: en el estudio del calibrado, el tamaño muestral de los grupos toma el valor $n_1 = n_2 = 10$, a la vez que $n_3 = 80$. Tras simular obtenemos el valor $\delta = 0{,}047$, cercano al nivel de significación $\alpha = 0{,}05$. Podemos deducir de esto que el test está bien calibrado.

2. $\boldsymbol{\gamma \neq 0}$: en este caso, estudiaremos diferentes valores de γ y veremos como se comporta la potencia del test al aumentar el propio γ y el tamaño muestral n. El tamaño muestral del grupo 1 y dos será $n_1 = n_2 = n/8$, mientras que el tamaño muestral del tercer grupo será $n_3 = n$.

Como podemos apreciar en la Figura 3.1, el test tiene un buen comportamiento en términos de potencia, ya que cuanto mayor es el tamaño muestral n más rápido converge a 1. Además, también se puede observar que cuanto mayor es $|\gamma|$ más rápido alcanzamos

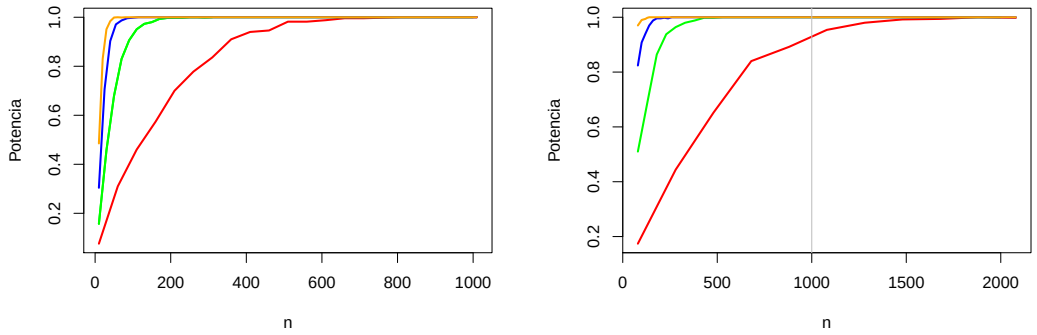

Figura 3.1: Gráficas de la potencia del test en función del tamaño muestral en el contraste de no efecto de la variable continua. Corresponde al Primer Escenario (gráfica izquierda) y al Segundo Escenario (gráfica derecha). Línea roja: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}05$. Línea verde: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}1$. Línea azul: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}15$. Línea naranja: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}2$.

una potencia próxima a 1.

Por último, al comparar las gráficas de la Figura 3.1[3], parece que el test no pierde potencia cuando las muestras están desbalanceadas. Así, en el Primer Escenario, cuando $n = 625$ (tenemos 1875 datos totales), la potencia del test toma un valor próximo a 1. De igual forma, en el Segundo Escenario, cuando $n = 1500$ (nuevamente 1875 datos totales), la potencia del test vuelve a estar próxima a 1. Podemos concluir que no se aprecian diferencias significativas en la potencia del test cuando las muestras están desbalanceadas.

Tercer escenario: diseño heterocedástico balanceado

A diferencia del Primer Escenario, ahora no se cumple una de las hipótesis del modelo, pues los datos no son homocedásticos. De esta forma, las desviaciones típicas del primer y segundo grupo son $\sigma_1 = \sigma_2 = 0{,}5$, mientras que la desviación típica del tercer grupo es $\sigma_3 = 2$.

1. $\boldsymbol{\gamma = 0}$: la simulación nos devuelve $\delta = 0{,}05$, igual al α fijado. Parece que pese al incumplimiento de la homocedasticidad el test tiene un buen comportamiento en

[3]Como en el Capítulo 2, se ha trazado una linea cuando $n = 1000$ con el objetivo de comparar el Primer Escenario (A1) con los demás escenarios.

términos de calibrado.

2. $\boldsymbol{\gamma \neq 0}$: como en el Primer Escenario, estudiaremos diferentes valores de γ y veremos como se comporta la potencia del test al aumentar el propio γ y el tamaño muestral n.

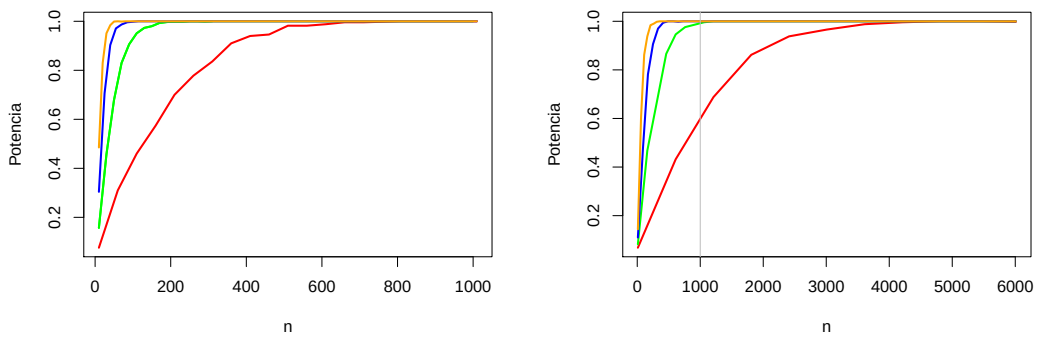

Figura 3.2: Gráficas de la potencia del test en función del tamaño muestral en el contraste de no efecto de la variable continua. Corresponde al Primer Escenario (gráfica izquierda) y al Tercer Escenario (gráfica derecha). Línea roja: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}05$. Línea verde: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}1$. Línea azul: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}15$. Línea naranja: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}2$.

De forma similar al anterior escenario, en la imagen izquierda de la Figura 3.2 podemos apreciar que cuanto más grande es el valor absoluto de la pendiente de regresión $|\gamma|$ y mayor es el tamaño muestral n mayor es la potencia del test. Además, el test sigue presentando un buen comportamiento en términos de potencia, pues a mayor tamaño muestral, más se aproxima la potencia del test a 1.

Sin embargo, si comparamos el Primer Escenario (A1) con el Tercer Escenario (A3), podemos ver que el test es más potente cuando los datos son homocedásticos que cuando son heterocedásticos. Así, cuando la pendiente de la recta de regresión toma el valor $\gamma = 0{,}05$ y los datos son homocedásticos, el test alcanza una potencia próxima a 1 con $n = 500$ observaciones en cada grupo. Por el contrario, cuando las poblaciones no comparten la misma desviación típica, necesitamos cerca de $n = 3500$ observaciones para alcanzar una potencia próxima a 1. Esto pone de manifiesto el papel que juega la homocedasticidad

de los datos en el comportamiento del test, y en consecuencia, en la información que nos aportan los datos.

Cuarto escenario: diseño heterocedástico desbalanceado

Por último, estudiaremos un escenario en el cual no se cumple la hipótesis de homocedasticidad y las muestras están desbalanceadas. Así $\sigma_1 = \sigma_2 = 0{,}5$ mientras que $\sigma_3 = 2$.

1. $\boldsymbol{\gamma = 0}$: la simulación nos devuelve $\delta = 0{,}046$, cercano al α fijado, por lo que pese a las carencias de los datos el test sigue estando bien calibrado.

2. $\boldsymbol{\gamma \neq 0}$: estudiaremos diferentes valores de γ y veremos como se comporta la potencia del test al aumentar el propio γ y el tamaño muestral n. Los grupos 1 y 2 serán de tamaño $n_1 = n_2 = n/8$, mientras que $n_3 = n$.

Como en los anteriores escenarios, en la Figura 3.3 podemos apreciar que el test tiene un buen comportamiento en términos de calibrado. En este caso parece más adecuado comparar el Tercer y Cuarto escenario, para así evaluar como afecta el desbalanceado de los grupos cuando los datos no son homocedásticos. Así, al comparar las gráficas del Tercer y Cuarto Escenario, podemos observar como aumenta el tamaño muestral necesario al desbalancear las muestras. De esta forma, en el Tercer Escenario, si $n = 5000$ (disponemos de 15000 datos), la potencia del test toma un valor próximo a 1. Por el contrario, en el Cuarto Escenario, tomando un $n = 12000$ (volvemos a tener 15000 datos), la potencia del test no toma un valor tan próximo a 1.

De esta forma, podemos ver que si se cumplen las hipótesis del modelo, que las muestras estén desbalanceadas no supone ningún problema en el contraste de no efecto de la variable continua. Sin embargo, al desbalancear las muestras, el desbalanceado toma importancia y supone una reducción en la potencia del test.

3.5.2. Contraste de no efecto de la variable discreta

Ahora la variable a contrastar es la variable discreta, debido a ello, fijaremos el valor de la pendiente de la recta de regresión $\gamma = 0$ y tomaremos diferentes valores de las medias de los grupos, μ_1, μ_2 y μ_3.

Primer escenario: diseño homocedástico balanceado

Como el Primer Escenario es análogo al del contraste de la variable continua, las desviaciones típicas de los grupos serán $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$.

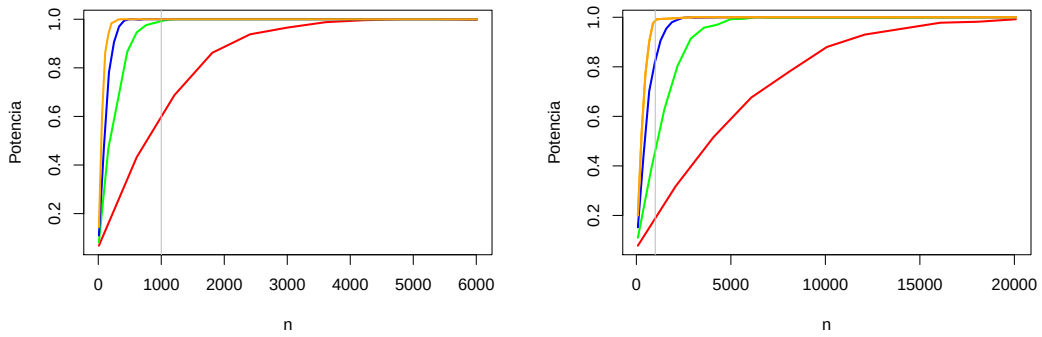

Figura 3.3: Gráficas de la potencia del test en función del tamaño muestral en el contraste de no efecto de la variable continua. Corresponde al Tercer Escenario (gráfica izquierda) y al Cuarto Escenario (gráfica derecha). Línea roja: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}05$. Línea verde: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}1$. Línea azul: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}15$. Línea naranja: la pendiente de la recta de regresión toma el valor $\gamma = 0{,}2$.

1. $\boldsymbol{\mu_1 = \mu_2 = \mu_3 = 1}$: tras simular obtenemos el valor $\delta = 0{,}051$, cercano al nivel de significación $\alpha = 0{,}05$. Podemos deducir de esto que el test está bien calibrado.

2. $\boldsymbol{\mu_1 = \mu_2 = 1}$ y $\boldsymbol{\mu_3 \neq 1}$: en este caso, estudiaremos diferentes valores de μ_3 y veremos como se comporta la potencia del test al aumentar el propio μ_3 y el tamaño muestral n.

En la Figura 3.4 podemos observar que cuanto mayor es la media del tercer grupo μ_3, más potente será el test. De igual forma, cuanto mayor sea el tamaño muestral, más potencia tendrá el test F. Debido a esto, podemos concluir que el test tiene un buen comportamiento en términos de potencia, ya que al aumentar el tamaño muestral, la potencia del test se aproxima a 1.

Segundo escenario: diseño homocedástico desbalanceado

Como mencionamos en el contraste de no efecto de la variable continua, en el Segundo Escenario las desviaciones típicas de los grupos serán $\sigma_1 = \sigma_2 = \sigma_3 = 0{,}5$ y las muestras estarán desbalanceadas.

1. $\boldsymbol{\mu_1 = \mu_2 = \mu_3 = 1}$: el tamaño muestral de los grupos será $n_1 = n_2 = 10$, mientras que $n_3 = 80$. Tras simular obtenemos el valor $\delta = 0{,}039$, cercano al nivel de

significación $\alpha = 0{,}05$. Podemos deducir de esto que el test está bien calibrado.

2. $\boldsymbol{\mu_1 = \mu_2 = 1}$ y $\boldsymbol{\mu_3 \neq 1}$: en este caso, estudiaremos diferentes valores de μ_3 y veremos como se comporta la potencia del test al aumentar el propio μ_3 y el tamaño muestral n. Los grupos 1 y 2 tomarán el valor $n_1 = n_2 = n/8$, mientras que el grupo 3 será $n_3 = n$.

El test, como podemos observar en la Figura 3.4, vuelve a mostrar un comportamiento bueno en términos de potencia, ya que esta aumenta según aumenta el tamaño muestral. De igual forma, podemos apreciar que apenas hay diferencias entre el Primer Escenario y el Segundo Escenario en lo que se refiere a la potencia del test respecto al tamaño muestral. Podemos concluir que no se aprecian diferencias significativas en la potencia del test cuando las muestras están desbalanceadas.

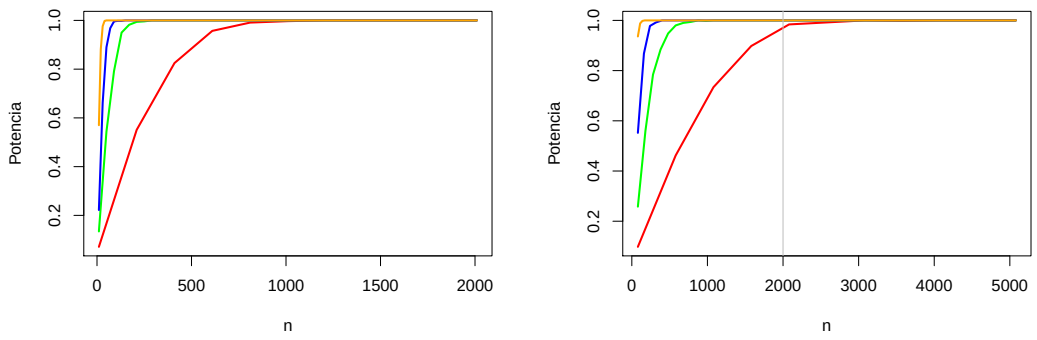

Figura 3.4: Gráficas de la potencia del test en función del tamaño muestral en el contraste de no efecto de la variable discreta. Corresponde al Primer Escenario (gráfica izquierda) y al Segundo Escenario (gráfica derecha). Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

Tercer escenario: diseño heterocedástico balanceado

En este escenario, las desviaciones típicas del primer y segundo grupo son $\sigma_1 = \sigma_2 = 0{,}5$, mientras que la desviación típica del tercer grupo es $\sigma_3 = 2$, por lo que los datos son heterocedásticos.

1. $\boldsymbol{\mu_1 = \mu_2 = \mu_3 = 1}$: tras simular obtenemos el valor $\delta = 0{,}085$, lejano al nivel de significación $\alpha = 0{,}05$. Podemos decir que el test no está bien calibrado.

2. $\boldsymbol{\mu_1 = \mu_2 = 1}$ y $\boldsymbol{\mu_3 \neq 1}$: en este caso, estudiaremos diferentes valores de μ_3 y veremos como se comporta la potencia del test al aumentar el propio μ_3 y el tamaño muestral n.

En la Figura 3.5 podemos apreciar que cuanto mayor es la media del tercer grupo μ_3 y mayor es el tamaño muestral n mayor es la potencia del test. El test presenta un buen comportamiento en términos de potencia, pues a mayor tamaño muestral, más se aproxima la potencia del test a 1.

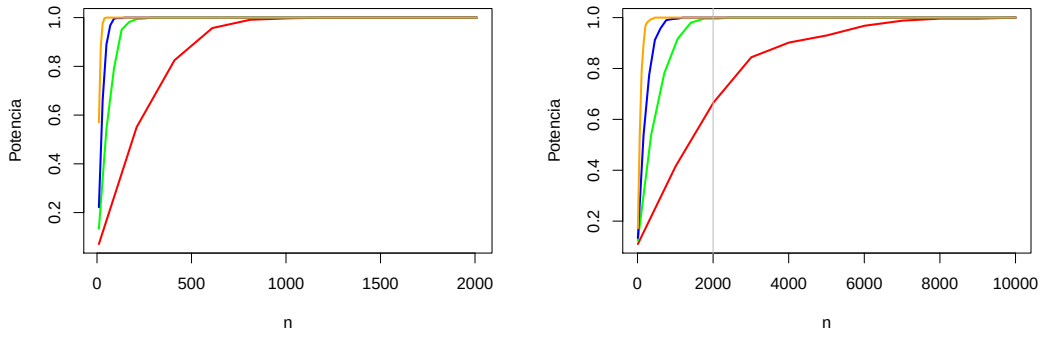

Figura 3.5: Gráficas de la potencia del test en función del tamaño muestral en el contraste de no efecto de la variable discreta. Corresponde al Primer Escenario (gráfica izquierda) y al Tercer Escenario (gráfica derecha). Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

Al comparar las gráficas de la Figura 3.5, se hace patente que el test es más potente cuando los datos son homocedásticos que cuando son heterocedásticos. De esta forma, cuando la media del tercer grupo toma el valor $\mu_3 = 1{,}5$ y los datos son homocedásticos, el test alcanza una potencia próxima a 1 con $n = 40$ observaciones en cada grupo. Sin embargo, si los datos no son homocedásticos, necesitamos cerca de $n = 250$ observaciones para alcanzar una potencia próxima a 1.

Cuarto escenario: diseño heterocedástico desbalanceado

Por último, evaluaremos que ocurre cuando las muestras están desbalanceadas y los datos son heterocedásticos. Las desviaciones típicas tomarán los valores $\sigma_1 = \sigma_2 = 0{,}5$ mientras que $\sigma_3 = 2$.

1. $\boldsymbol{\mu_1 = \mu_2 = \mu_3 = 1}$: en el estudio del calibrado tomaremos un tamaño muestral de $n_1 = n_2 = 10$, a la vez que $n_3 = 80$. Tras simular obtenemos el valor $\delta = 0$, lejano al nivel de significación $\alpha = 0{,}05$. Podemos decir que a pesar de que el test no rechace nunca H_0, no está bien calibrado.

2. $\boldsymbol{\mu_1 = \mu_2 = 1}$ y $\boldsymbol{\mu_3 \neq 1}$: estudiaremos diferentes valores de μ_3 y veremos como se comporta la potencia del test al aumentar el propio μ_3 y el tamaño muestral n. Los grupos 1 y 2 serán de tamaño $n_1 = n_2 = n/8$, mientras que $n_3 = n$.

En la Figura 3.6 se puede observar el comportamiento de la potencia respecto al tamaño muestral en el Cuarto Escenario (A4). Está claro que el test tiene un buen comportamiento en términos de potencia ya que converge a 1, sin embargo, como en el contraste de la variable continua, debido al desbalanceado de los datos es necesario un mayor tamaño muestral.

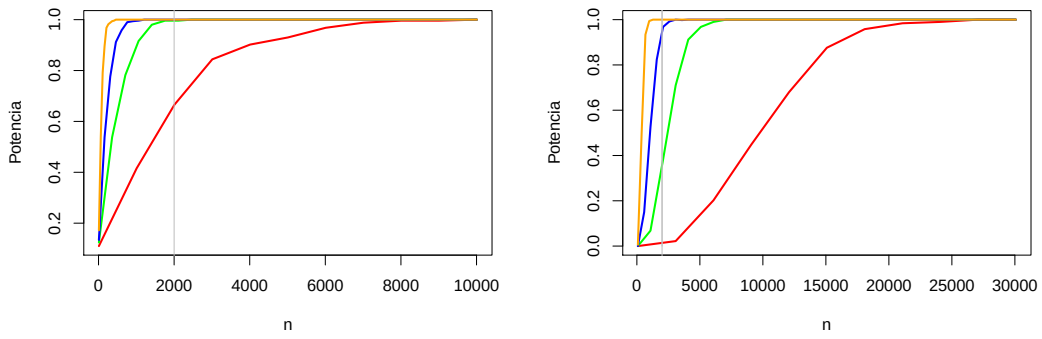

Figura 3.6: Gráficas de la potencia del test en función del tamaño muestral en el contraste de no efecto de la variable discreta. Corresponde al Tercer Escenario (gráfica izquierda) y al Cuarto Escenario (gráfica derecha). Línea roja: la media del tercer grupo toma el valor $\mu_3 = 1{,}1$. Línea verde: la media del tercer grupo toma el valor $\mu_3 = 1{,}2$. Línea azul: la media del tercer grupo toma el valor $\mu_3 = 1{,}3$. Línea naranja: la media del tercer grupo toma el valor $\mu_3 = 1{,}5$.

3.6. Conclusiones

Tras realizar el estudio de simulación en términos de calibrado y potencia de los contrastes de la variable discreta y de la variable continua en el modelo ANCOVA, podemos observar:

- **Datos desbalanceados**: como en el caso del modelo ANOVA, es recomendable usar muestras balanceadas. En caso de que las muestras estén desbalanceadas, si los datos cumplen las hipótesis del modelo ANCOVA, obtendremos buenos resultados en términos de calibrado y potencia. Sin embargo, si no se cumple la hipótesis de homocedasticidad, el test es menos potente si los datos no están balanceados (esto lo podemos ver en el Cuarto Escenario).

- **Datos heterocedásticos**: tenemos dos situaciones diferentes:

  - Datos balanceados: en caso de que las muestras sean balanceadas, los test tendrán un comportamiento aceptable en términos de calibrado. Por otra parte, ambos test tendrán un peor comportamiento en términos de potencia, necesitando un tamaño muestral mayor para alcanzar una potencia próxima a 1.
  - Datos desbalanceados: el contraste de la variable continua parece tener un buen comportamiento en términos de calibrado. En el de la variable discreta, si los grupos de menor tamaño presentan una mayor desviación típica, habrá una mayor cantidad de falsos positivos, mientras que si los grupos de mayor tamaño tienen una mayor desviación típica, aumentaran los falsos negativos (esto se puede apreciar en el Cuarto Escenario en la Figura 3.6). En el estudio de la potencia, debemos de tener en cuenta las desviaciones de los grupos y la varianza de los errores, ya que de ello dependerá el comportamiento de los test.

  Podemos comprobar esta hipótesis mediante el test de Levene.

- **Tamaño muestral**: cuanto mayor sea el número de datos, mayor será la potencia de los test, por lo que más fiables serán nuestros resultados.

Bibliografía

Corder, G. W. y Foreman, D. I. (2011). Nonparametric statistics for non-statisticians.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

Maxwell, S. E., Delaney, H. D., y Kelley, K. (2017). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Routledge.

McCullagh, P. (2018). *Generalized Linear Models*. Routledge.

Moore, D. S. (2005). *Estadística Aplicada Básica*. Antoni Bosch editor.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wu, S., Jin, Z., Wei, X., Gao, Q., Lu, J., Ma, X., Wu, C., He, Q., Wu, M., Wang, R., et al. (2011). Misuse of statistical methods in 10 leading chinese medical journals in 1998 and 2008. *The Scientific World Journal*, 11.